RICE UNIVERSITY

Essays in Efficiency Analysis

by

Pavlo Demchuk

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Robin C. Sickles, Chair
Reginald Henry Hargrove Chair of
Economics,
Rice University

Peter R. Hartley,
George and Cynthia Mitchell Chair in
Sustainable Development and
Environmental Economics,
Rice University

David W. Scott,
Noah Harding Professor of Statistics,
Rice University

HOUSTON, TEXAS
August 2012

ABSTRACT

Essays in Efficiency Analysis

by

Pavlo Demchuk

Today a standard procedure to analyze the impact of environmental factors on productive efficiency of a decision making unit is to use a two stage approach, where first one estimates the efficiency and then uses regression techniques to explain the variation of efficiency between different units. It is argued that the abovementioned method may produce doubtful results which may distort the truth data represent. In order to introduce economic intuition and to mitigate the problem of omitted variables we introduce the matching procedure which is to be used before the efficiency analysis. We believe that by having comparable decision making units we implicitly control for the environmental factors at the same time cleaning the sample of outliers. The main goal of the first part of the thesis is to compare a procedure including matching prior to efficiency analysis with straightforward two stage procedure without matching as well as an alternative of conditional efficiency frontier. We conduct our study using a Monte Carlo simulation with different model specifications and despite the reduced sample which may create some complications in the computational stage we strongly agree with a notion of economic meaningfulness of the newly obtained results. We also compare the results obtained by the new method with ones previously produced by Demchuk and Zelenyuk (2009) who compare efficiencies of Ukrainian regions and find some differences between the two approaches.

Second part deals with an empirical study of electricity generating power plants before and after market reform in Texas. We compare private, public and municipal power generators using the method introduced in part one. We find that municipal power plants operate mostly inefficiently, while private and public are very close in their production patterns. The new method allows us to compare decision making units from different groups, which may have different objective schemes and productive incentives. Despite the fact that at a certain point after the reform private generators opted not to provide their data to the regulator we were able to construct tree different data samples comprising two and three groups of generators and analyze their production/efficiency patterns.

In the third chapter we propose a semiparametric approach with shape constrains which is consistent with monotonicity and concavity constraints. Penalized splines are used to maintain the shape constrained via nonlinear transformations of spline basis expansions. The large sample properties, an effective algorithm and method of smoothing parameter selection are presented in the paper. Monte Carlo simulations and empirical examples demonstrate the finite sample performance and the usefulness of the proposed method.

Acknowledgments

Culmination of graduate work – a dissertation, written by one, supervised by a few and read by a dozen before going public. It carries the knowledge of many generations of scientists and, as it happens in economics, from different fields. Intended to some extent "push the frontier", analyze historical events and provide results not available before, as well as find answers or possibly pose new questions dissertation provides a summary of research, co-operation and co-ordination abilities of a student.

My own work during the stay at Rice University and this thesis in particular would not be possible without people and events which either contributed or in other ways helped me to carry it out. On and off the "academic field" I have received a lot of ideas, support, criticism and disapproval which I believe helped me to optimally conduct my work given the constraints.

Professor Ximing Wu, who is the main author of the third chapter, provided a great intellectual challenge for me. Although my part of the paper was an empirical application it still required the understanding of the material, which is not very straightforward. I thank professor Wu for letting me be the part of this interesting project.

A major role in my academic life was played by professor Robin Sickles. Having met him in Kyiv in 2006 I did not imagine how powerful and important that meeting would be in, at the time, near future. Providing advice in the process of my development as a scientist, giving me part in the organization of two amazing North American Productivity Workshops, editing my work and having me referee articles for Empirical Economics, Journal of Econometrics and Journal of Productivity Analysis, teaching us

topics in statistics oftentimes neglected by textbooks, are just a few things I am grateful for. And the list goes on. Thank You, Robin!

I am grateful to my professors for teaching me how to think, tackle academic and nonacademic problems, cope with time, stress and sleep deprivation. Let me assure you, it was not in vain. Very big "Thanks!" goes to Professor David Scott for teaching great statistics courses and being there ready to help and answer our questions, even at 2 a.m.

I also thank Professor Hartley for giving me useful advice during school and over the course of the writing of my thesis. I am grateful for ideas provided by Professor Jay Zarnikau which helped me with the "electricity" chapter of this work. Participants of Productivity workshops in the US and Europe are to be acknowledged for their feedback on my "work in progress" which evolved into this thesis, I presented in New York 2008, Pisa 2009 and Verona 2011.

Special gratitude goes to the Ukrainian community of Houston and especially the Dub and the Dijak families, who supported me and my family in particularly complicated situations as well as everyday life. You have taught me who real Ukrainians away from Ukraine are. I admire your energy, enthusiasm and dedication despite all the hardships you had and have to go through. Дуже Вам дякую![1]

No words can describe all the gratitude to my family. My mother, father and brother who unconditionally supported my aspirations at the same time criticizing me for spending too much time away from home. I thank my wife Natalya for patiently handling

[1] Thank you very much! (Ukr.)

my regular visits and especially departures every half a year, for her strength in taking care of our son Ivan, who recently grew impatient and keeps asking when am I coming home, and our daughter Mariya, whom I haven't met yet. I am grateful for the inspiration when I needed it the most. For those moments of joy which gave me the will and strength to carry on. Thank you for everything!

Contents

List of Figures

List of Tables

Chapter 1

On a Set of Methods to Address Problems with Two-Stage DEA

1.1. Introduction

Oftentimes when analyzing productive efficiency in a sample of decision making units (DMUs) the researcher is as interested in the impact of external factors on levels of efficiency, as in the levels of efficiency themselves. The external factors do not enter the production function directly, but may possibly influence efficiency[2]. For example, one may think of a situation in which comparisons among several subgroups within the sample are the main concern or situations in which differences among DMUs before and after an "epochal event" (e.g. financial crisis) are of primary interest. This sort of analysis may be carried out in the framework of efficiency score estimation first and then a subsequent second stage, wherein the bias corrected efficiency scores are regressed

[2]Also known as external factors, non-discretionary factors or environmental factors.

against an indicator variable[3]. If a formal statistical model is posited then the statistical integrity of any two-stage procedure requires that the environmental variables used in the second stage should be independent of the input variables used to estimate technical efficiency in the first stage. This is easily done in a regression setting but less easily implemented when using linear programming approaches. A number of methods to deal with external factors were developed since the original paper of Banker and Morey (1986), who first suggested a model that incorporates the influence of nonproductive factors. These methods can be divided into two main categories: one-stage and multiple-stage approaches.

One-stage approaches incorporate external, or as they are sometimes called, "non-discretionary" or "environmental" factors, into the model via additional constraints. Banker and Morey (1986) treat these factors as internal inputs and do not optimize the efficiency over them. By conducting simulated analysis Ruggiero (1996) shows that the above approach leads to biased estimation, due in part to the infeasibility of the frontier given the level of external inputs that the DMU faces. To overcome this problem he sets up a model where DMUs in more favorable environment (e.g. socio-economic status of families on the student achievement study and the use of the term "environmental" factors) can be excluded from the reference set. Unfortunately the model breaks down when the number of external inputs exceeds unity. Moreover, excluding observations from the model leads to biased estimates of efficiency since the new reference sets are

[3] We do not take up the other obvious issue that the treatment in this example may be endogenously chosen and thus that the lack of random assignment causes additional problems in interpretation as has been pointed out in the extensive treatment effects literature.

smaller and more units will be defining the frontier and thus be fully efficient. Syrjänen (2004) proposes a generalized model which incorporates not only external and internal inputs, but also distinguishes between volume and index type factors. Syrjänen's model contains parameters that are usually used to control for undesirable outputs to distinguish between external/internal factors/inputs and outputs. Despite interesting suggestions no formal tests for such separability are provided. Yang and Paradi (2006) developed a method to handicap inputs and outputs. In their method they penalize DMUs in a more advantageous environment by employing a higher input-handicapping measure and/or a lower output handicapping measure that increases inputs and/or decreases outputs of specific DMUs while at the same time assigning disadvantaged units a lower input-handicapping measure and/or larger output-handicapping measure. The most recent attempt was made by Löber and Staat (2010). In their model, authors add a constraint to the DEA problem that excludes certain observations from the reference set (dummy variable restriction). The constraint contains indicators with zeros for inputs to be excluded from the set and ones for the inputs to be included. In general, one stage methods add constraints with non-discretionary variables to the DEA model but mainly disregard their influence on productive efficiency. If one assumes that no relationship exists between output and external variables, as well as a single frontier, while in reality output and non-discretionary variables are correlated, then one would think that efficiency scores will be biased due to the omitted variable bias that would be evident in a regression-type setting. Moreover, if there are indeed different groups within the sample they may operate under slightly different technologies, and therefore evaluating both groups using the same frontier will lead to the underestimation of efficiencies.

Once efficiency scores are obtained further evaluation is often conducted by multistage models. Ruggiero (1998) provides an alternative to his earlier model by estimating the DEA model without external factors in the first stage and regressing efficiency scores on the external factors expected to influence efficiency. He then constructs an index based on the obtained results and estimates another DEA model using the abovementioned index. Muñiz (2002) suggested a three stage model with slacks. He considers the DEA model without external variables with slack variables for inputs and outputs at the first stage. Once slacks are obtained one evaluates a DEA model for the slack of each variable. After quantifying slacks during the second stage the original slack values from the first stage are decomposed into two components – the influence of external inputs and true technical inefficiency. Next, original data is adjusted using slacks obtained in the second stage and thus the third stage slack values, as argued, provide pure inefficiency effects. The bias corrected two-stage method provided by Simar and Wilson (2007) is aimed at avoiding the bias problem encountered in the first stage of efficiency score estimation due to the unobservability of the true production frontier. Therefore, at the first stage DEA scores are estimated using inputs and outputs that directly enter the production function. Then at the second stage truncated regression with bootstrap is proposed in order to describe the influence of the external factors on the efficiency of the decision making units. Their model requires two main assumptions: the separability of production stage and external factors, as well as the functional form specification in the second stage, which leaves some researchers skeptical towards the approach. Conditional nonparametric frontier defined probabilistically was considered by Daraio and Simar (2007) who estimate efficiency scores by setting up a one-stage conditional DEA model,

where the distribution of (x, y) pairs is conditioned on the external factors z. Measures based on the Daraio and Simar m-frontier are estimated using a multi-stage approach. Furthermore the influence of the external factors is evaluated. Using information from the conditional and unconditional efficiency frontier, the authors find the direction of influence of one external factor. For the case of multiple z's Daraio and Simar (2007) point out that it would be hard to obtain any information on the marginal effect of external factors on efficiency if the z's are correlated. Two stage procedures are quite popular because they are easy to interpret and communicate the results to a variety of consumers of such methods, such as policy makers, regulators, and businesspeople. Such methods have been utilized in a large number of studies, among the more recent of which are the analysis of Slovenian farms (Gocht and Balcombe 2006), Ukrainian regions (Demchuk and Zelenyuk 2009), and the Greek prepared meat products industry (Keramidou et al. 2010).

The abovementioned multistage methods do not regard external factors as part of the production function or do not optimize the production set over these factors similarly to one stage models. Therefore if external factors are correlated with output $E(y \mid x, z) \neq 0$, but are disregarded in the first stage it would seem natural that an omitted variable bias would be introduced that would lead to biased efficiency scores[4], thus making any inference unreliable. Bias due to the omitted variables not accounted for during the efficiency estimation in the SF setting, considered by e.g. Caudill and Ford (1993), Ruggiero (1996), Wang and Schmidt (2002) is found by the means of Monte

[4] For more details see Wang and Schmidt (2002).

Carlo simulation and was not described formally. In the two-stage DEA setting such bias was never formally considered, mainly due to the peculiarities of the setup of the first stage linear optimization and second stage statistical models and the impossibility to integrate the two. Therefore, the idea of the omitted variable bias exists in the literature, but so far has been treated only via Monte Carlo simulations.

The alternative approach to the solution of the biasedness problem in the DEA framework is addressed in this paper by pretreating observations within the sample before conducting efficiency score estimation. We consider division of the sample into two distinct groups (similar to the introduction of the dummy variable in the parametric case; multiple groups will be considered later) based on external variables that do not directly enter the production function. There exist a number of potentially good methods which may provide some insights on the importance of the external factors on the production process and efficiency, for example: propensity score matching, discriminant analysis, and principal component analysis but we focus on propensity score matching methods to account for external factors. We also conduct Monte Carlo experiments on artificially generated data and compare the full sample without matching and two subsamples – one with matched data and one without.

The remainder of the paper is structured as follows. Section 2 describes the general production model and then modifies it to account for external factors. We also introduce a propensity score matching procedure prior to the DEA estimation in order to control for external factors. Section 3 contains Monte Carlo simulation results of the comparison of the performance of the two stage efficiency analysis with and without propensity score matching. In Section 4 we move to an empirical setting and analyze

factors that influence changes in efficiency of Ukrainian regions using our propensity score matching methods. These results are then compared with those obtained by Demchuk and Zelenyuk (2009). Section 5 concludes.

1.2. Production Model

1.2.1. General Formulation

Usually, efficiency analysis is carried out by estimating a production possibility frontier (PPF) and relating DMUs to this frontier via a distance measure. Estimation may take a nonparametric (Data Envelopment Analysis - DEA) or a parametric (Stochastic Frontier Analysis - SFA) form. The distance between each particular observation and the frontier is summarized by the efficiency score which indicates how efficient the DMU is relative to others and to the firms and their convex combinations that define the frontier.

Decision making units ($j = \overline{1,J}$) engaged in the production process use input(s) $x \in \Re_+^k$ to produce output(s) $y \in \Re_+^m$. The technology used to transform inputs into outputs is assumed to be accessible to all DMUs and can be characterized by the technology set T:

$$T = \{(x, y) \in \Re_+^k \times \Re_+^m \mid x \ can \ \ produce \ \ y\} \tag{1}$$

We assume that technology can be characterized by a set of standard regularity conditions from production theory:

a) A1. The technology set T is closed and non-empty.

b) A2. $(x, 0_m) \in T, \ \forall x \in \Re_+^k$. It is possible to waste resources and produce nothing.

c) A3. $(0_k, y) \notin T$. It is impossible to produce any output without using inputs. This assumption is also known as a "no free lunch" assumption.

d) A4. $P(x) \equiv \{y : (x, y) \in T\}$ is bounded $\forall x \in \Re_+^k$.

e) A5. $P(x) \neq (0_m)$ for some $x \in \Re_+^k$. Technology is productive.

f) A6. $(x, y) \in T \Rightarrow (x', y') \in T, \ \forall x' \geq x, \ y' \leq y$. Inputs and outputs are freely disposable.

Given these assumptions on the technology set, the smallest convex cone that contains all the data can be characterized by:

$$T \equiv \{(x, y) : \sum_{j=1}^{J} \lambda_j y^j \geq y \geq 0_m, \ \sum_{j=1}^{J} \lambda_j x^j \leq x, \ \lambda_j \geq 0, \ \ j = \overline{1, J}\} \qquad (2)$$

Here the variable $\lambda_j \geq 0$ is the intensity level of activity j. The smallest convex cone that contains all the data for the constant returns to scale assumption (CRS) technology we consider in this paper is the solution to the optimization problem:

$$\begin{aligned} &\max_{\theta, \lambda} \theta_i \\ &\sum_{j=1}^{J} \lambda_j y_{mj} - \theta_i y_{mi} \geq 0, \ \ m = \overline{1, M} \\ &x_{ki} - \sum_{j=1}^{J} \lambda_j x_{kj} \geq 0, \ \ k = \overline{1, K} \\ &\lambda_j \geq 0, \ \ j = \overline{1, J} \end{aligned} \qquad (3)$$

In this work we consider output orientation of the production model, i.e. output maximization given fixed resources. The approach is easily extended to the input orientation, i.e. minimization of resources given a fixed output level. Under the CRS

assumption the results from both models are equivalent. Given the technology, one may estimate the efficiency score defined by Debreu-Farrell as:

$$\delta(x,y) \equiv \max_{\theta}\{\theta : (x,\theta y) \in T\} \tag{4}$$

In the output oriented case $\delta(x,y) \in [1,\infty)$ and an efficiency score of 1 indicates that the DMU is fully efficient. When $\delta(x,y) > 1$ the DMU is operating inefficiently and the percentage of inefficiency is $\left[1 - \frac{1}{\delta(x,y)}\right] \times 100\%$.

Of course the true technology is never observed and the true (potential) output level is not known. Thus the inefficiencies need to be estimated using the data at hand via the Data Envelopment (DEA) estimator:

$$\hat{T} \equiv \{(x,y) : \sum_{j=1}^{J} \lambda_j y^j \geq y \geq 0_m,\ \sum_{j=1}^{J} \lambda_j x^j \leq x,\ \lambda_j \geq 0,\ \ j = \overline{1,J}\} \tag{5}$$

Consistency of the DEA estimator $\hat{T}$ has been shown in Kneip et al. (1998). The observed inefficiency measure $\hat{\delta}(x,y)$ is a downward biased estimator due to the fact that the estimated technology set $\hat{T}$ is a subset of the true T. Despite that, $\hat{\delta}$ is a consistent and asymptotically unbiased estimator of the true inefficiency δ as is shown in (Kneip et al. 1998).

1.2.2. Environmental Factors

We extend the general model with an introduction of the environmental factors. Each DMU uses input(s) $x \in \Re_+^k$ to produce output(s) $y \in \Re_+^m$. External to the explicit

production process exist environmental factors $z = \Re^r$, which do not influence the production process directly, but impact efficiency of each particular DMU. Now the triple (x, y, z) defines the technology set, as opposed to only (x, y) in the previous definition of T.

Simar and Wilson (2007) add new assumptions that govern the production process accounting for environmental factors, which we use in this study as well. These are:

g) B1. The sample observations (x_i, y_i, z_i) are realizations of iid random variables with pdf $f(x, y, z)$ and support over $T \times \Re^r$, where $T \subset \Re_+^{k+m}$ is a technology set defined in (1). Also, $f(x, y, z) \neq f(x, y)$.

The joint distribution $f(x, y, z)$ can be described as a series of conditional distributions: $f(x_i, y_i, \delta_i, z_i) = f(z_i) f(\delta_i \mid z_i) f(x_i, y_i \mid \delta_i, z_i)$.

h) B2. The conditioning in $f(\delta_i \mid z_i)$ from the previous equation is assumed to be operating though the following mechanism: $\delta_i = g(z_i, \beta) + \varepsilon_i \geq 1$, where $g(\cdot)$ is a smooth, continuous function, β is a vector of parameters and ε_i is a continuous iid random variable independent of z_i.

i) B3. ε_i is distributed $N(0, \sigma_\varepsilon)$ with left truncation $1 - g(z_i, \beta)$ for each i.

j) B4. For all $(x, y) \in T$ such that $(\theta^{-1}x, y) \notin T$ for $\theta > 1$, $f(x, y \mid z)$ is strictly positive and continuous in any direction toward the interior of T (the frontier) for $\forall z$.

k) B5. $\forall (x, y)$ in the interior of T, $\delta(x, y \mid T)$ is differentiable in both its arguments.

Here, assumptions B1-B3 were introduced in order to account for the environmental variables. In particular B1 and B2 are needed to ensure the separability condition, meaning that the external variable z_i influences the production process only via δ_i. Assumptions B.2 and B.3 in particular are crucial and quite restrictive assumptions in that they assume the independence of z's from the level of the benchmark's firms, i.e., those that define the frontier, and rather are variables that effect the relative efficiency of a firm. Moreover, the randomness of the relative efficiency scores δ_i is uncorrelated with the z's. Thus any statistical noise that moves a firm below the efficient frontier must be independent of any proxies one has to control for factors that are meant to determine why a firm is below the efficient frontier. This appears at first blush to be an unrealistic an untenable assumption. Assumptions B4-B5 used by Simar and Wilson (2007) are extensions of earlier work by Kneip et al. (1998) and are fairly standard regularity conditions.

1.3. Stage Zero: Propensity Score Matching

Given a set of observations one may be interested in particular subsets and their performance. We consider two alternatives – estimate a PPF using the whole dataset and compare the groups of interest, or estimate two separate frontiers[5]. To estimate a single frontier one needs to make the assumption that the same technology is utilized by all

[5] We assume that grouping is done using a specific rule that may be done manually, e.g. DMUs in the 75th percentile of output, large vs. small DMUs, we call this rule the "Ad Hoc Rule". Another way to group DMUs is according to some "epochal event", which helps us to differentiate the "before" and "after" state of the DMUs, or by way of a grouping based on regulatory oversight, such as governmental versus private firms in the same industry.

DMUs or at least is accessible by different subsets of DMUs. If this assumption fails then in each of these two cases some DMUs will be compared to a production frontier that differs from the frontier consistent with the technology under which the DMU operates, leading us to biased results. By estimating two different technology frontiers we can compare units within each group, but are unable to immediately compare units between groups.

Furthermore, a researcher may be interested in the impact of variables that do not enter the production function, but possibly have an impact on efficiency. Researchers either add constraints to the DEA model and/or regress efficiency score on these variables. Such procedures will provide biased results, since the DEA model is misspecified due to the omitted variables.

In order to mitigate the biasedness problem described above we suggest a method that is used in the treatment-effects literature propensity score matching (PSM). The only paper published to this date, namely, "Technology Adoption and Technical Efficiency: Organic and Conventional Dairy Farms in the United States" written by C.Mayen, J. Balagtas and C. Alexander in the American Journal of Agricultural Economics (2010), where authors use PSM prior to estimating the stochastic frontier. In contrast to our work these authors use matching to address self-selection to different technologies and are not interested in the environmental factors per se. Out of twenty observed factors used in the first stage – matching, only eight are used to explain inefficiency in the stochastic frontier setting.

The propensity score is typically used to evaluate the effects of treatment (medical, educational, economic, historical, e.g. some epochal event, etc.), on a specific group of individuals or DMUs as compared to the group of same DMUs that have not been treated.

In our analysis we use the generally accepted notation, where treatment is denoted by $\mathcal{T}$, with $\mathcal{T}=1$ if treated and $\mathcal{T}=0$ otherwise[6]. Responses to treatment and no treatment Y_1 and Y_0 respectively are used together with pre-treatment variables Z and treatment $\mathcal{T}$ to "sample" the observations under study.

The majority of studies concentrate on the so-called Average Treatment Effect on the Treated (ATT) which represents the cause of treatment on the treated DMU.

The average treatment effect on the treated is obtained from the following expression:

$$\gamma_{ATT} = E(Y_{1i} - Y_{0i} \mid Z_i, \mathcal{T}_i = 1) = E(Y_{i1} \mid Z_i, \mathcal{T}_i = 1) - E(Y_{i0} \mid Z_i, \mathcal{T}_i = 1) \tag{6}$$

Another effect sought after in treatment effect studies is the Average Treatment Effect (ATE):

$$\gamma_{ATE} = E(Y_{i1} - Y_{i0}) = E(Y_{i1} \mid \mathcal{T}_i = 1) - E(Y_{i0} \mid \mathcal{T}_i = 0) \tag{7}$$

Although we are not interested in the treatment effects per se, note that the estimation of the average treatment effect on the treated is a comparison of the nearest DMUs under different production frontiers, given all the covariates, while the average

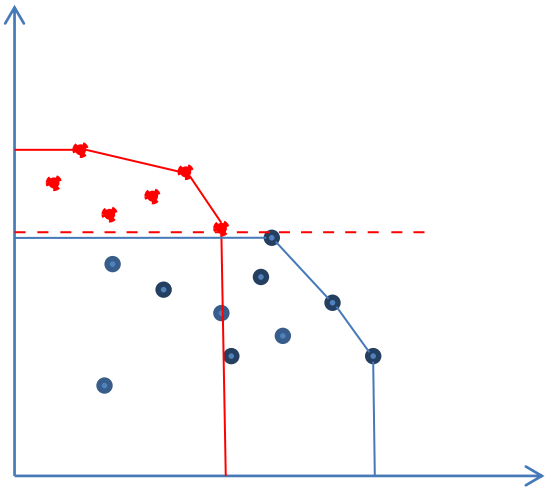
[6] Here we consider a binary treatment, although we can extend the discussion to multiple treatments, e.g. Lechner 2001, and thus consider multiple matching approaches in the two-stage DEA model.

treatment effect is a difference between the nearest DMUs under different frontiers without covariates. Usually $ATT \neq ATE$, but in the DEA framework it is possible that both effects coincide, because the ad hoc rule produces similar results to the propensity score matching procedure. This may be due to a small dataset or well defined ad hoc rule.

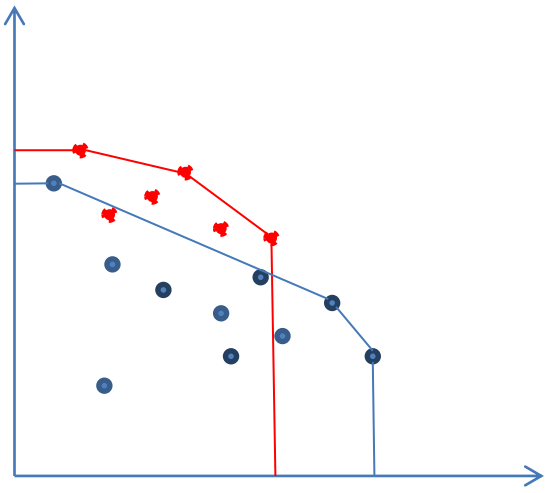
The expression for $E(Y_{i1} \mid Z_i, \mathcal{T}_i = 1)$ is easily obtained from the data at hand, while $E(Y_{i0} \mid Z_i, \mathcal{T}_i = 1)$ needs to be estimated, since we do not observe the counterfactual mean of the DMU had it not been treated. Simply utilizing $E(Y_{i0} \mid Z_i, \mathcal{T}_i = 0)$, the mean of the untreated units to approximate $E(Y_{i0} \mid Z_i, \mathcal{T}_i = 1)$ will result in a selection bias. The propensity score $P(Z_i)$ matching method uses all the information available assuming independence of Y_{i0} and Y_{i1} as well as treatment conditional on external factors to eliminate the bias [$Bias = E(Y_{i0} \mid Z_i, \mathcal{T}_i = 1) - E(Y_{i0} \mid Z_i, \mathcal{T}_i = 0)$]. Propensity score matching allows us to compare averages of the treated and untreated DMUs by eliminating the bias:

$$E(Y_{i0} \mid P(Z_i), \mathcal{T}_i = 1) = E(Y_{i0} \mid P(Z_i), \mathcal{T}_i = 0) = E(Y_{i0} \mid P(Z_i))$$

Figure 1. Ad Hoc Rule vs. Propensity Score Matched Subsamples



a) AH Rule



b) PS Matched

The advantages of the proposed application of matching before employing efficiency analysis over the straightforward ad hoc rule (AH) has several benefits. Consider as an example, DMUs producing output Y using input X. Our goal is to compare units within the sample based on the size of their output (e.g. top 75% of output

(Large) vs. the rest (Medium and Small)). We may proceed in several ways: 1) assuming that all units operate under the same technology with possibly different returns to scale (we evaluate a DEA with constant returns to scale); 2) assuming that some external factors distinguish one group from another we rank DMUs by output level, choose a quartile (e.g., the 3rd quartile (above the dashed dotted line in Figure 1 a)) and estimate two different frontiers; 3) assuming different influence of external factors on DMUs between the groups and using all the information on inputs and output at hand we match observations based on their propensity score, the probability of the unit belonging to a certain sub-group (Figure 1 b)). This idea can be generalized to any problem with indicator variable(s).

As we mentioned above and also referring to Wang and Schmidt (2002), ignoring the fact that efficiency is dependent on external factors in the estimation of the single frontier will give us biased results. The second approach, or the AH rule will also give us biased results, as seen from Figure 1, since it does not take external factors into the account either.

The approach we are suggesting, propensity score matching, explicitly takes into account all the external factors that may be available to the researcher, thus reducing the omitted variable bias in efficiency estimation. We should note that it is possible for the groups identified from the AH rule to coincide with groups produced by PSM, which may be due to uninformative external factors used for matching or well devised AH. Despite the fact of possible coincidence PSM would appear to be a better approach, since it always accounts for external factors. Moreover, a correct AH is problematic when more

than one output is available, while PSM can be employed in an environment with multiple outputs/treatments (e.g. good output - value added and bad output - pollution).

Having identified groups of DMUs one may proceed with two-stage estimation of the influence of external environmental factors on efficiency in each group. As a corrected efficiency score evaluation procedure we suggest the following algorithm:

Stage 0: Preliminary Stage. Estimate the probability of each DMU belonging to the group of interest (propensity score) via a probability model (e.g. logistic regression). Divide the sample at hand into several groups (i.e. matched and unmatched) by matching the propensity scores.

Stage 1: Efficiency Estimation. Estimate frontiers for each group separately, where frontiers are based only on factors that enter the production function directly.

Stage 2: Estimation of the Influence of External Factors on Efficiency. Estimate using bias correcting methods, e.g. truncated regression with bootstrap (Simar and Wilson (2007)) of efficiency scores obtained in Stage 1 on the external factors that do not enter the production function directly.

To formalize the algorithm we describe its procedure in the remaining part of this section.

Stage 0: Preliminary Stage.

[1]. Estimate binary logistic regression to find the probability of receiving treatment:

$$P(\mathcal{T}_i | Z_i = z_i) = E(\mathcal{T}_i) = \frac{e^{z_i\beta_i}}{1+e^{z_i\beta_i}} = \frac{1}{1+e^{-z_i\beta_i}} \quad (8)$$

Using the logit function, equation (8) transforms into a generalized linear model:

$$\log[P(\mathcal{T}_i)/1-P(\mathcal{T}_i)] = z_i\beta_i \quad (9)$$

Model (9) is estimated with the maximum likelihood and propensity scores are obtained by plugging respective $\hat{\beta}_i$'s into (9).

[2]. Match the propensity scores for treated and control participants P_i and P_j, respectively, by nearest neighbor technique:

$$C(P_i) = \min_j \left\| P_i - P_j \right\|, \quad i \in I_1, j \in I_0 \quad (10)$$

Where $C(P_i)$ is the neighborhood which contains a control participant j (from the set of control participants I_0 as a match for the treated participant i (from the set of treated participants I_1), if the absolute difference between propensity scores is the smallest among all possible pairs of propensity scores between i and j. Once j is matched to i, j is removed from I_0 without replacement.

Stage 1 and 2: Efficiency Estimation and External Factors Influence (based on Simar and Wilson, 2007)

[3]. Using the matched sub-sample obtained in [2] compute efficiency scores $\hat{\delta}_i = \hat{\delta}(x_i, y_i \mid \hat{\wp}), \ \forall i = \overline{1,n}$

[4]. Use maximum likelihood to obtain estimates $\hat{\beta}_i$ of β_i and an estimate $\hat{\sigma}_\varepsilon$ of σ_ε in the truncated regression of $\hat{\delta}_i$ on z_i (use $m < n$ if $\hat{\delta}_i > 1$).

[5]. Obtain n sets of bootstrap estimates of efficiency scores $\mathfrak{B}_i = \{\hat{\delta}_{ib}^*\}_{b=1}^{L_1}$, where L_1 is a number of bootstrap iterations.

[6]. For each $i = \overline{1,n}$ compute the bias corrected scores $\hat{\hat{\delta}}_i = \hat{\delta}_i - Bias(\hat{\delta}_i)$ using the bootstrap estimates $\mathfrak{B}_i$ and the original estimates $\hat{\delta}_i$.

[7]. Use maximum likelihood to obtain estimates $\hat{\hat{\beta}}_i$ and $\hat{\hat{\sigma}}_\varepsilon$ from the truncated regression of $\hat{\hat{\delta}}_i$ on z_i.

[8]. Obtain a set of bootstrap estimates $\mathfrak{G} = \left\{\left(\hat{\hat{\beta}}_i^*, \hat{\hat{\sigma}}_\varepsilon^*\right)_b\right\}_{b=1}^{L_2}$, where L_2 is a number of bootstrap iterations.

[9]. Construct confidence intervals for each element in β_i and σ_ε using bootstrap values in $\mathfrak{G}$ and the original estimates of $\hat{\hat{\delta}}_i$ and $\hat{\hat{\sigma}}_\varepsilon$.

1.4. Monte Carlo Simulation

In this section we pursue a number of sampling experiments to assess the potential for our suggested methods vis-à-vis those that are currently in use. It is important to keep in mind therefore that when examining the usefulness of particular approaches to two-stage estimation, a data generating process such as that used in Simar

and Wilson (2007), which involves a set of external factors that influence efficiency directly but have no direct impact on production, i.e., one that assumes separability of external factors and output, will not in general be an appropriate estimator. We explain how to include the external influence on production implicitly by modifying the DGP setup originally suggested by Simar and Wilson (2007).

In general, for each observation i, external factors are specified as follows: we set $z_{i1}=1$ and randomly choose $z_{ij}\sim N(\mu_z,\sigma_z^2)$ for $j=\overline{2,r}$. Next, we generate a left truncated[7] error term which will be used in the regression: $\varepsilon_i\sim N(0,\sigma_\varepsilon^2)$, where $\varepsilon_i=1-z_i\beta$. Then set the inefficiency to be $\delta_i=z_i\beta+\varepsilon_i$. Inputs are: $x_{ij}\sim U(5,20)$ for $j=\overline{1,p}$. Total output is specified by the following relationship: $y_i=\delta_i^{-1}\sum_{j=1}^{p}x_{ij}^k$. If more than one output is needed, then total output y_i is split according to shares defined as $\alpha_1\sim U(0,1),\quad \alpha_j\sim U\left(0,1-\sum_{j=1}^{q-2}\alpha_j\right)$ for $j=\overline{2,q-1}$. Therefore $y_{ij}=\alpha_j\delta_i^{-1}\sum_{j=1}^{p}x_{ij}^k$ for $j=\overline{1,q-1}$ and $y_{iq}=\left(1-\sum_{j=1}^{q-1}\alpha_j\right)\delta_i^{-1}\sum_{j=1}^{p}x_{ij}^k$.In the case of multiple outputs we must be very careful. When generating multiple outputs we can't simply split one output into several ones using only one α, because this way treatments on both outputs will be the same.

---

[7] Simar and Wilson (2007) suggest a modified transformation method. Let $\Phi(\bullet)$ and $\Phi^{-1}(\bullet)$ denote standard normal distribution and standard normal quantile functions respectively, so that $u=\Phi^{-1}(\Phi(u))$. Generate $v\sim U(0,1)$, let the adjusted truncation point be $c'=c/\sigma$, and set $v'=\Phi(c')+[1-\Phi(c')]v$, then compute $u=\sigma\cdot\Phi^{-1}(v')$ to get the left truncated normal deviate.

Therefore by matching on the first treatment, second treatment or both the same subsample will be created. To avoid this problem we need to generate a different α for each observation.

For our experiments we set the number of external factors to be two and four $(r = 2,4)$ including the constant, $\mu_z = 2$, $\sigma_z = 2$, $\sigma_\varepsilon = 1$, $k = 3/4$. Moreover, coefficients for β's need to be specified, for simplicity we assume $\beta_1 = \beta_2 = 0.5$. Correlation between x_1 and z_1 exists in our model, with $\rho(x_1, z_1) = (0.2, 0.5, 0.8)$.

The core idea of matching, after obtaining the estimated propensity scores, is to create a new sample of cases that share approximately similar likelihoods of being assigned to the treatment condition. A possible drawback of the method is the restriction of the final sample to only matched observations. We examine several scenarios and evaluate the proposed approach based on root-mean-square errors of parameter estimators and "coverage intervals", i.e. the proportion of the total number of Monte Carlo experiments where the confidence interval covers the true value of the β's and σ_ε. In each scenario three samples are evaluated: the full sample, as well as reduced matched and reduced unmatched samples. Our results show very little difference between matched and unmatched subsamples; therefore we will describe the results of the matched data only.

We consider the case of one output, two inputs and one external factor ($r = 2$). Alas, the majority of applied studies do not consider such a simple relationship, i.e.

dependence of efficiency on one external factor. Therefore we decided to drop this model from the paper[8] and consider a more reasonable case with multiple external factors.

We simulated 100 Monte Carlo trials is each case. In the bootstrap procedure for the first loop used to obtain bias-corrected efficiency measures we had $B_1 = 500$ replications, and $B_2 = 1000$ for the second loop in which truncated regression was estimated. Each case is considered with $n = 100$ and $n = 400$ observations. Note that after one to one matching (top 25% of DMUs, by the level of production) we reduce samples to $n = 50$ and $n = 200$ units per sample, respectively.

Table 1. Estimated Confidence Intervals Coverage in Truncated Regression (Full Model)

Model	Full								
	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
n=100									
β_0	0.36	0.55	0.92	0.36	0.54	0.86	0.36	0.63	0.96
β_1	0.96	0.99	0.99	0.89	0.95	0.97	0.85	0.90	0.97
β_2	0.89	0.91	0.97	0.90	0.93	1.00	0.90	0.96	0.98
β_3	0.92	0.94	0.98	0.93	0.95	0.99	0.96	0.96	0.97
σ_ε	0.96	1.00	1.00	0.89	1.00	1.00	0.85	1.00	1.00
n=400									
β_0	0.31	0.42	0.64	0.34	0.45	0.73	0.40	0.54	0.77
β_1	1.00	1.00	1.00	0.98	0.99	1.00	0.94	0.98	0.99
β_2	0.96	0.98	0.99	0.96	0.98	1.00	0.99	1.00	1.00
β_3	0.99	0.99	1.00	0.98	0.98	0.99	0.95	0.98	1.00
σ_ε	1.00	1.00	1.00	0.98	1.00	1.00	0.94	1.00	1.00

[8]These results are available upon request.

Case 1. Correlated Model. Two outputs, three inputs, and three external factors with binary treatment on producing in the top quartile and $\rho(x_1, z_1) = (0.2, 0.5, 0.8)$.

Tables 1 and 2 report results for the full and matched models, two different sample sizes and three different levels of correlation between x_1 and z_1. They both provide results comparable with the ones found by Simar and Wilson (2007) and are plausible on both accounts.

The full model exhibits low coverage of constant term β_0 in smaller confidence bounds, while the matched model provides better coverage for the intercept. Similar to Simar and Wilson (2007) we observe that with the increase of the sample size as well as the confidence interval size, coverage level increases within each model. When comparing between the full and matched models coverages of all variables with the exception of β_0 as mentioned above, are similar in magnitude.

Table 2. Estimated Confidence Intervals Coverage in Truncated Regression (Matched Subset)

Model	Matched								
	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
n=50									
β_0	0.77	0.94	0.99	0.80	0.97	1.00	0.83	0.93	0.99
β_1	0.98	1.00	1.00	0.89	0.94	0.97	0.94	0.95	0.96
β_2	0.92	0.92	0.96	0.87	0.92	0.94	0.94	0.94	0.97
β_3	0.95	0.97	0.98	0.93	0.96	0.97	0.89	0.91	0.93
σ_ε	0.98	1.00	1.00	0.89	1.00	1.00	0.94	1.00	1.00
n=200									
β_0	0.70	0.90	0.98	0.71	0.86	1.00	0.62	0.85	0.99
β_1	0.96	0.98	0.99	1.00	1.00	1.00	0.99	1.00	1.00
β_2	0.95	0.95	0.99	0.98	0.99	1.00	0.92	0.96	0.97
β_3	0.96	0.98	1.00	0.92	0.93	0.98	0.94	0.95	0.97
σ_ε	0.96	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00

The next step of our analysis is the comparison of root-mean-square errors (RMSE) of the estimators produced by full and matched models. Results presented in Table 3 show that RMSEs are on average smaller with the larger sample at hand, with the exception of some instances of σ_ε in cases of small and medium correlation. This means that bias correction in smaller samples increases RMSE relative to the larger samples. Recall that we observed better coverages for the intercept in the matched model, which may be the result of wider confidence intervals.

Table 3. Root-Mean-Square Error of Parameter Estimators (Full vs. Matched Models)

Model	Full			Matched		
	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
n=100/50						
β_0	0.3088	1.2521	1.7059	1.7557	1.5389	7.8626
β_1	0.0325	0.0903	0.1452	0.1897	0.0714	1.3927
β_2	0.0278	0.0786	0.1898	0.2273	0.3046	1.3502
β_3	0.1355	0.1544	0.1006	0.0462	0.1515	0.5001
σ_ε	0.1282	0.5217	0.5312	0.2940	0.1945	2.1051
n=400/200						
β_0	0.1902	0.3501	0.5549	0.2859	0.3550	0.7704
β_1	0.0763	0.0281	0.1330	0.0354	0.0636	0.1655
β_2	0.0080	0.0235	0.0280	0.0686	0.0395	0.0309
β_3	0.0154	0.0277	0.0389	0.1483	0.0029	0.0123
σ_ε	0.3877	0.3072	0.5270	0.6590	0.4255	0.8990

Comparing the full and matched model we see that the results differ. Between the levels of correlation on average matched model has higher RMSEs, which is extremely pronounced in the case with high correlation between x_1 and z_1. With a larger sample, coefficients that are not correlated with inputs of the production model have smaller RMSEs in the matched model as compared to the full model. Also note that the slight

superiority of the full model is likely due to the sample size, since we have observed that bias correction performs better in larger samples.

Case 2.Omitted Variables. We extend Case 1, by assuming that one of the production factors (x_2) was omitted in the efficiency estimation stage, but was considered in the second stage as an explanatory variable in the truncated regression. From Tables 4 and 5 we notice that in the full and matched models confidence intervals almost completely cover all the coefficients with the only exception of σ_ε that is covered only by the 95 and 99% confidence intervals in the matched model and 99% confidence interval in the full model. This result means that the estimated distribution of σ_ε is far from the true value of 1 and only the tail of the distribution covers the true value.

Table 4. Estimated Confidence Intervals Coverage in Truncated Regression with First Stage Omitted Variable (Full Model)

Model	Full								
	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
n=100									
β_0	0.95	0.97	1.00	0.97	0.99	1.00	0.96	0.99	1.00
β_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
β_2	1.00	1.00	1.00	0.98	0.99	1.00	0.99	0.99	1.00
β_3	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
σ_ε	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00
n=400									
β_0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
β_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
β_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
β_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
σ_ε	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00

An interesting result is obtained from the matched model (Table 5), where the coefficient of the correlated variable z_1 is not covered by the confidence interval (only in

small sample, 1 to 3 percent in the case with $\rho = 0.2$. Besides this, wide confidence intervals cover the true value of β's in all specifications and do it more precisely that confidence intervals from Case 1.

Table 5. Estimated Confidence Intervals Coverage in Truncated Regression with First Stage Omitted Variable (Matched Subset)

Model	Matched								
	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
n=50									
β_0	0.96	0.99	1.00	0.98	1.00	1.00	0.98	1.00	1.00
β_1	0.01	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.04
β_2	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
β_3	0.97	0.97	0.99	0.93	0.97	0.99	0.97	0.98	0.99
σ_ε	0.01	0.98	0.99	0.00	0.99	1.00	0.00	0.98	0.99
n=200									
β_0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
β_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
σ_ε	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00

Table 6 contains RMSE values for full and matched models. We again observe that models with smaller sample size have higher root-mean-square errors. Matched model contains smaller RMSEs for the most of $\rho = 0.2$ case coefficients as well for $\rho = 0.5$, despite the smaller sample. For the high correlation case the full model performs better.

In summary, truncated regression with bootstrap with or without matching performs better with the larger sample size dataset. Bias correction in smaller samples increases RMSE relative to the larger samples. With a large number of external variables confidence intervals become wide, as compared to the previous findings of Simar and

Wilson (2007). In the model of omitted variable we found that the external factor correlated with one of the inputs became insignificant in the regression setting. It is hard to give preference to any of the two models, since both perform similarly to each other. Therefore, it is safe to conclude that a matching procedure may be carried out prior to efficiency estimation and truncated regression in order to better assess the influence of environmental variables on efficiency of DMUs.

Table 6. Root-Mean-Square Error of Parameter Estimators with First Stage Omitted Variable (Full vs. Matched Models)

Model	Full			Matched		
	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
n=100/50						
β_0	0.4853	1.7350	0.5689	0.1399	0.7825	3.5365
β_1	0.0518	0.4491	0.4658	0.1799	0.2122	1.3790
β_2	0.0775	0.5265	0.2555	0.2739	0.3792	1.3736
β_3	0.5976	0.6577	0.6528	0.1556	0.1217	0.9445
σ_ε	0.5584	3.7495	2.8764	0.4003	1.4771	3.7751
n=400/200						
β_0	1.3280	1.1706	1.2355	1.4052	1.8101	0.2073
β_1	0.1213	0.2342	0.1939	0.3001	0.2345	0.5857
β_2	0.0788	0.2401	0.2134	0.1624	0.2556	0.2266
β_3	0.6366	0.6590	0.6559	0.0374	0.1940	0.2034
σ_ε	1.8236	2.5132	3.3221	2.1094	2.5656	5.1733

Despite the fact that after matching the sample size is reduced the procedure itself is sounder theoretically, since it accounts for the external factors prior to any estimation mitigating omitted variable bias. Moreover, for a large sample size the proposed procedure provides results comparable to the truth.

An interesting extension to the current work is to consider multiple treatments, where two cases are of particular interest: multiple treatments in cross section (e.g.

different types of pollutants - CO_2 (1 or 0), NO_x (1 or 0), etc.) and in single variable (e.g. types of fertilizer – fertilizer 1, 2, 3, etc.).

1.5. Conditional Efficiency Simulation

An approach accounting for external factors via the conditional measure originally proposed for the robust frontiers by Cazals et al. (2002) and extended to full frontiers by Daraio and Simar (2005, 2007) provides an alternative way to evaluating effects of external factors on productive efficiency. This procedure considers a conditional full frontier:

$$S_{Y|X,Z}(y|x,z) = Prob(Y \geq y|X \leq x, Z = z) \tag{11}$$

relative to which the output conditional efficiency is measured:

$$\lambda(x,y|z) = \sup\{\lambda|S_{Y|X,Z}(\lambda y|x,z) > 0\} \tag{12}$$

A smoothed estimator of the full frontier with compact support kernel K and appropriate bandwidth h_n is obtained via:

$$\hat{S}_{Y|X,Z,n}(y|x,z) = \frac{\sum_{i=1}^{n} I(X_i \leq x, Y_i \geq y) K((z-Z_i)/h_n)}{\sum_{i=1}^{n} I(X_i \leq x) K((z-Z_i)/h_n)} \tag{13}$$

with efficiency measure:

$$\hat{\lambda}_n(x,y|z) = \sup\{\lambda|\hat{S}_{Y|X,Z,n}(\lambda y|x,z) > 0\} =$$

$$\max_{\{i|X_i \geq x, |Z_i - z| \leq h\}} \left\{ \min_{j=1,\ldots,q} \left(\frac{Y_i^j}{y^j} \right) \right\} \tag{14}$$

The evaluation of external factors global effect is obtained using the smoothed nonparametric regression framework. Regression for output orientation is written as:

$$Q_i^z = g(Z_i) + \epsilon_i, \quad i = \overline{1,n}, \tag{15}$$

where $Q_i^z = \frac{\hat{\lambda}_n(X_i,Y_i|Z_i)}{\hat{\lambda}_n(X_i,Y_i)}$, ϵ_i is an error term independent of external factors and $g(\cdot)$ is the mean regression function which may be estimated using the smoothed nonparametric regression introduced by Nadaraya (1964) and Watson (1964):

$$\hat{g}(z) = \frac{\sum_{i=1}^{n} K\left(\frac{z-Z_i}{h}\right) Q_i^z}{\sum_{i=1}^{n} K\left(\frac{z-Z_i}{h}\right)} \tag{16}$$

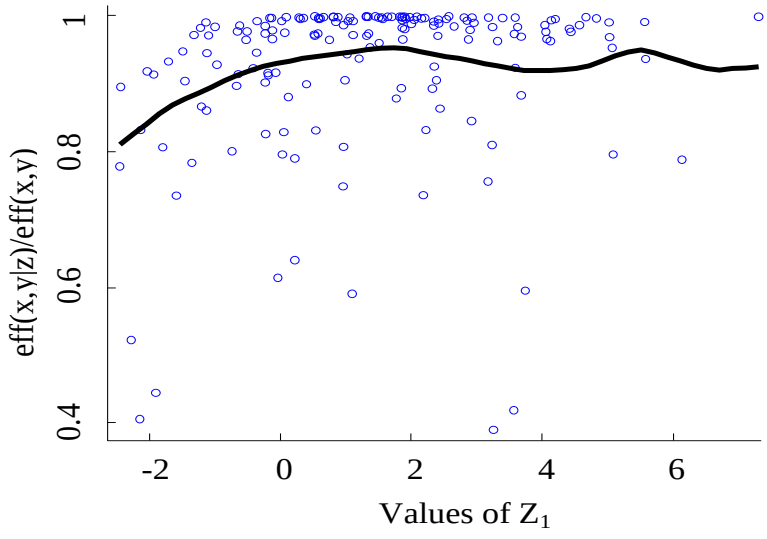
In practice, after finding optimal bandwidths one fixes the reference data set over which efficiency scores are computed, furthermore in the regression setting data points corresponding to the case $Q_i^z = 1$ are dropped as they do not provide any information about Z_i influencing efficiency.

Results of such a model may be interpreted qualitatively but the size of each effect remains unknown since it varies at different levels of Z_i. In the output case, increasing regression is interpreted to have a negative effect of the external factor on productive efficiency, while decreasing regression will correspond to a positive effect. Note, that for a model with single external variable – a case rarely considered in practice, solution is straightforward. Determination of optimal bandwidths in case of multiple external factors using a product kernel of n univariate kernel functions suggested by Bădin, et al. (2010) as one of the approaches, possesses a feature similar to the matching method proposed in this paper, namely, as pointed put by Hall et al. (2004) irrelevant

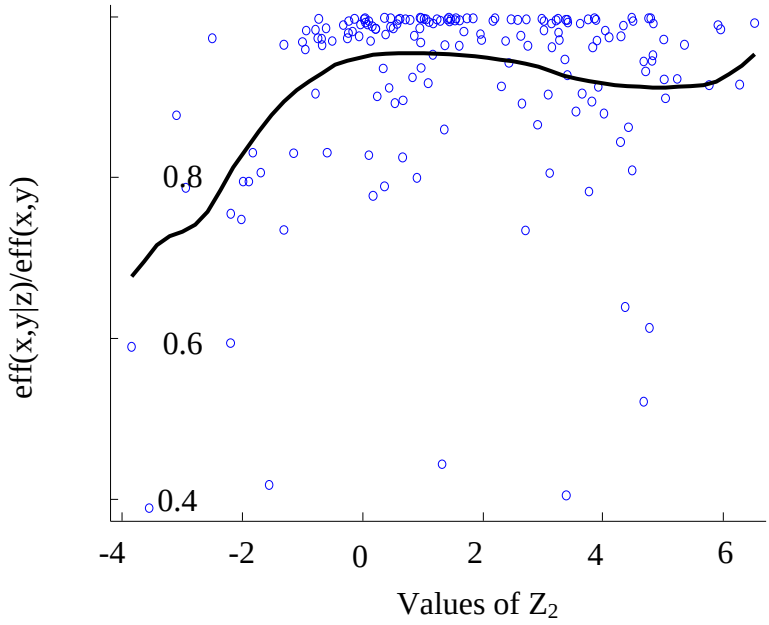
components of Z will be oversmoothed thus reducing the sample by dropping irrelevant observations. Increased dimensionality of the model and possible cross-correlation between the external factors may severely impede the recovery of true marginal effects. Moreover, in applied work researchers usually deal with samples containing between fifty to several hundred observations that has models with even 2 inputs, 2 outputs and 2 external factors - space of 6 dimensions, produce results plagued by the "curse of dimensionality" and makes their interpretation very unreliable.

Next we will demonstrate a simulated example of the conditional efficiency approach. We use the data generated according to the DGP used earlier in this paper and consider a two output, three inputs and three external factors model with (ρ=0.8) and a constant returns to scale model.

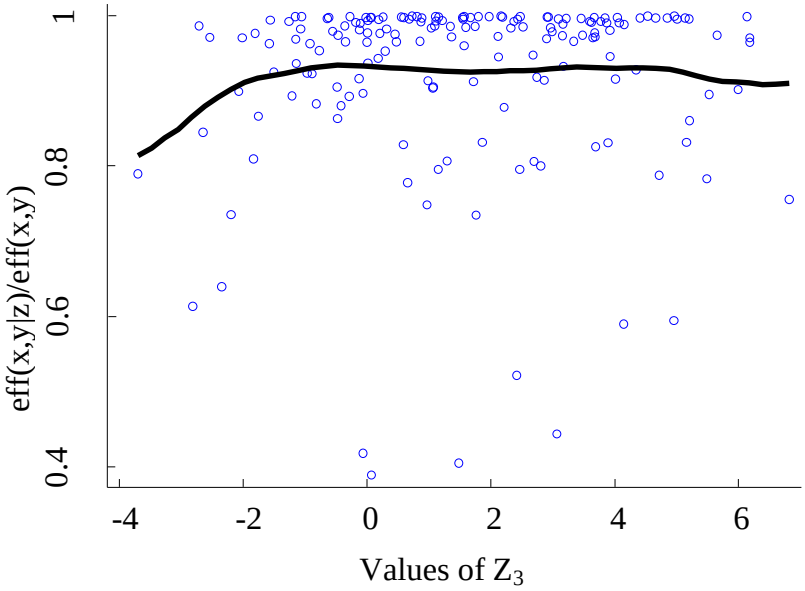
Figure 2. Effects of a Multivariate Z on a Full Frontier



a)



b)



c)

From the above figure we observe that according to the conditional efficiency approach Z_1 and Z_3 have similar effects: negative values of these external factors have a

negative effect on efficiency, while positive values correspond to a minor positive effect with some variability in Z_1. Second external variable appears to have more pronounced effects on efficiency: similarly to other two factors negative values have negative effects, then we observe positive effect of Z_2 between $(0.5; 5)$ and negative effect of larger positive values.

Conditional efficiency method captures some localized positive effects of external factors, but the overall effects are shown to be either neutral or negative, which is contrary to the DGP used here. Such a result is not surprising as Dario and Simar (2007) noted that a case with correlated external factors will be hard to evaluate due to the convolution issues.

1.6. Empirical Example

In this section we provide an empirical illustration of the use of our procedure, by reexamining the analysis of Demchuk and Zelenyuk (2009). Demchuk and Zelenyuk (2009) test productive efficiency of Ukrainian regions after the economic stabilization in 1996. Gross value added (GVA) was their only output, while labor and capital were considered as inputs. At the second stage, variables which possibly influence the efficiency, such as the amount of foreign direct investment per worker, capital per worker, gross value added per worker, alcohol and tobacco consumption per capita and crime rate, were considered. In this application we will use information on the regional economy of Ukraine during the period 1999-2002. According to the National Bank of Ukraine, the highest level of economic growth of 9.2%, between 1999 and 2002 was recorded during 2001. Therefore we choose this year to be our “epochal event” and use it

as a treatment. Results from the truncated regression are presented in Table 7. Models 3 through 6 correspond to same models in Demchuk and Zelenyuk (2009). All coefficients are significant at the 5% level with the only exception of GVA per worker and GVA per worker squared, which are insignificant when appear together in the model. All models show that alcohol and tobacco consumption, foreign direct investment and crime rate are negligible in their influence on efficiency in the region, as opposed to the results obtained previously.

Table 7. Truncated Regression with Matching. Ukrainian Regions, 1999-2002

Name	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Constant	-2.444	-2.552	-2.321	-2.143	-2.437	-2.389
Industrial	-0.075	-0.034	-0.079		-0.091	
East	-0.012	-0.080		0.101		0.012
log(AlcTob)	0.009	0.072	0.002	-0.030	-0.096	-0.025
log(FDI_w)	0.006	-0.003	-0.007	0.022	0.002	0.0003
log(Cap_w)	3.332	3.344	3.219	3.307	3.322	3.327
log(GVA_w)	-6.143*	-3.166	-0.912*	-6.881*	-3.066	-3.183
log(GVA_w^2)	1.507*		-1.088*	1.838*		
log(Crim)	-0.006	0.004	0.032	-0.060	0.004	-0.018
σ_ε^2	0.018	0.016	0.019	0.015	0.018	0.018

* Insignificant at 5%

The effects of the indicator variables for a particular region having a predominately industrial base or being located in the eastern part of the country are also rather small and there is a negative influence on inefficiency if the region produces more industrial output and negligible influence if it is located in the east. Previously, these variables were found to negatively influence efficiency at 5% significance level. Capital per worker negatively influences efficiency, contrary to what was observed without our new PSM procedure. This may suggest that Ukrainian regions are overcapitalized. This observation is plausible, since before 1991 the economy was concentrated on the

production of capital goods. We should note that a large portion of obsolete capital remains on the balance sheets of companies in Ukraine. Gross value added per worker significantly adds to the efficiency in the region, indicating that the wealth of the region is positively correlated with the efficiency level, a result found earlier.

From the regression analysis and comparison of the results to previous estimations without matching we find some significant differences. Our results suggest that regions that were similar in their potential to those that experienced the highest growth in 1991-2001 succeeded because they produced more industrial output, while the obsolete capital remaining on the balance sheets was not an asset, but rather a detrimental factor to efficiency. Previously found negative influence of alcohol and tobacco consumption, foreign direct investment and crime rate proved to be negligible in regions with higher potential.

1.7. Conclusions

In this chapter we consider a problem that is oftentimes neglected by the researchers, namely the inclusion of external factors in the second step regression of efficiency analysis, while omitting them at the first stage. An omitted variable bias is carried over to the second stage and may undermine the veracity of study results. We propose an alternative approach using Propensity Score Matching to mitigate the problem. By selecting observations that are matched on all observable characteristics and may potentially be included in the second stage we are able to obtain results that are consistent with explicit conditioning. We show that this method may encounter some problems in small samples, but in large samples performs similarly to the conventional

truncated regression with bootstrap without matching. It was noted that bias correction in smaller samples increases RMSE relative to the larger samples. With large number of external variables confidence intervals become wider, as compared to the previous findings of Simar and Wilson (2007). In the model of omitted variables we found that the external factor correlated with one of the inputs became insignificant in the regression setting. We conclude that matching procedure may be carried out prior to efficiency estimation and truncated regression in order to better assess the influence of environmental variables on efficiency of DMUs. Comparing this method to an alternative – conditional efficiency approach we observe that correlated variables distort the result to an extent when the results are misleading. In addition to that we test the results in an empirical application and find that the matched sample brings us to somewhat different conclusions found previously. Despite the fact that after matching the sample size is reduced our procedure itself is sounder theoretically, since it accounts for the external factors prior to any estimation thus implicitly considering them in the analysis. Finally an interesting observation we were able to make analyzing results obtained by different approaches - they depend on the DGP under consideration, which may be an issue in itself. Since conventional methods in efficiency analysis do not allow for multiple treatments or may have very slow convergence rates if based on conditional distributions using kernel smoothers, we will be considering models with several treatments.

Chapter 2

Power Generating Utilities in Texas after Electricity Market Deregulation

2.1. Introduction

The wave of deregulation of state electricity markets in the 1990's and subsequent suspension or delay of reforms in some of them has motivated consumers, politicians and scientists to debate the success of this reform. Intended for competition in generation and retail, deregulation was expected to reduce electricity prices. Despite the expectations, competition retail prices had gone up in absolute terms, yet different analysts often report opposite findings. Axelrod, et al. (2006) by means of controlling for the effects of fuel costs conclude that wholesale electricity prices are lower after the reform. Apt (2005) and Joskow (2006) find no proof of lower prices, while Rose and Meeusen (2005) did not discern any overall benefits to consumers after deregulation. On the other hand, Zarnikau and Whitworth (2005) detect evidence of increases in electricity prices charged to residents in Texas, especially in the areas opened to competition and Zarnikau, et al. (2007) find similar patterns in commercial electricity prices. Some explanation of the absolute rise in electricity prices, especially in the deregulated areas, was sought in the

structure of fuels used for electricity generation. According to Rose (2007) evidence shows that fuel prices have played a role in escalating electricity prices, but the effects attributed to fuel prices and especially the so-called “marginal fuel”, which in many areas of North America is natural gas, are exaggerated. He hypothesizes that customer load and its seasonal variation seem to better explain price fluctuations.

Deregulation of the electricity market was, in part, implemented by the unbundling, or in some cases divestment of vertically integrated electricity companies into separate generation, transmission/distribution and retail entities. Generation facilities and retailers were allowed to compete for resources and customers, while the transmission and distribution sector remained regulated in recognition of its natural monopoly characteristics. Companies were allowed to consist of generation and retail units, but transmission facilities had to be operated separately.

Observing substantial increase in retail electricity prices, rising fuel prices and customer loads it is unclear how electricity market deregulation has influenced its efficiency. Did the openness to competition in generation and retail sectors improve productive efficiency or did long term contracts and excess capacity reduce it?

While the majority of studies concentrated on the rising prices, historical developments in the North American electricity market (e.g. Sioshansi and Pffafenberger, 2006), efficiency of market organization (e.g. Mansur and White, 2009), productive efficiency analysis of the generating market has been overlooked. Limited by the installed capacity, its age and nameplate efficiency power generators compete in an open

market providing cheapest electricity to its consumers, making productive efficiency instrumental in the possibility of winning the customer.

Before taking on a more ambitious project of analyzing changes in efficiency of the restructured electricity generation, transmission and retail sectors in the US we consider a smaller, but nonetheless important part of the issue – electricity generation market in Texas[9]. Considered to be one of the best restructured competitive markets in the world (e.g. RED index) Texas seems to be a great subject for analysis due to several advantages. First, the Electric Reliability Council of Texas (ERCOT) - an independent system operator (ISO) exclusively occupies nearly all the territory of Texas, does not overlap with any other state and remains mostly unconnected to any other state's grid. This feature allows ERCOT to operate only under the regulations of the state of Texas, unlike other ISO's that need to coordinate the actions of its constituents according to the laws adopted in different states where they operate. Second, electricity market is internally oriented, meaning that the largest share of electricity produced is consumed within Texas' borders. Since there is essentially no interstate electricity trading, under the Federal Power Act - ERCOT is not regulated by the Federal Energy Regulatory Commission (FERC), which makes it a unique entity.

[9] Several ideas for this project came from personal communication with Dr. Jay Zarnikau the president of Frontier Associates and formerly the Director of Electric Utility Regulation at the Public Utility Commission of Texas and a Program Manager at the UT College of Engineering's Center for Energy Studies.

Dominated by large vertically integrated investor-owned companies, Texas has seen a drastic increase in the number of participants since the introduction of the wholesale and retail competition in 1995 and 2002 respectively. By the unbundling of vertically integrated utilities regulators aimed at separating generators who operate with almost no economies of scale and transmitters who have large economies.

Today more than a quarter of total U.S. natural gas is produced in Texas and as a result almost half of state's electricity is generated at natural gas-fired power plants. Another 40% is produced using coal. The largest share of coal extracted and consumed in Texas is lignite coal, which is low in energy content but is also low on sulfur which makes it "friendlier" to the environment. Nonetheless, here consumption of coal is the biggest in the nation, which makes Texas one of the largest carbon dioxide (CO_2) and sulfur dioxide (SO_2) emitters. Nuclear and renewable electricity generating facilities are responsible for about 6% and 11% of total electricity generation by respective sources in the U.S., but in total state electricity generation account for only 10% and 2% respectively. Introduction of a competitive market in Texas and the increase of participants in electricity generation are followed by a decrease in pollution.

Considering the data available from the Federal Energy Regulatory Commission on power generators in Texas the impact of electricity market reform on productive efficiency of regulated and unregulated power generators is evaluated. The impact of competition in power generation on the improvements in efficiency of power plants is analyzed. We hypothesize that despite the creation of user friendly and competitive market, efficiency of regulated generators stayed approximately the same due to active

contracts and agreements, supply and demand constrains, as well as the inability and lack of incentive among generators to build enough capacity to compete with newer facilities.

To test our hypothesis we employ a method introduced in the previous chapter – preliminary propensity score analysis for space reduction followed by the nonparametric DEA method. Furthermore, to analyze the main drivers of productive efficiency of power generators after the reform we construct a Malmquist Index and consider its decomposition into efficiency and technical changes. For the estimation of efficiency we use installed capacity - proxy for capital, and average number of employees at a power plant as inputs and net generation as output. The data comes from the FERC website. To be more precise, we use the data on organic-fueled or combustible renewable steam-electric and gas turbine plants regardless of current ownership and/or operation. The time period under study is 1994-2003.

The remainder of the paper is structured as follows: an overview of the Texas electricity market and its reform is provided in Section 2; propensity score matching and data envelopment analysis methods are described in Section 3, together with the algorithm of Malmquist index decomposition; Section 4 contains description of the data sample, and in Section 5 we present estimated results. Section 6 concludes.

2.2. Texas Electricity Market

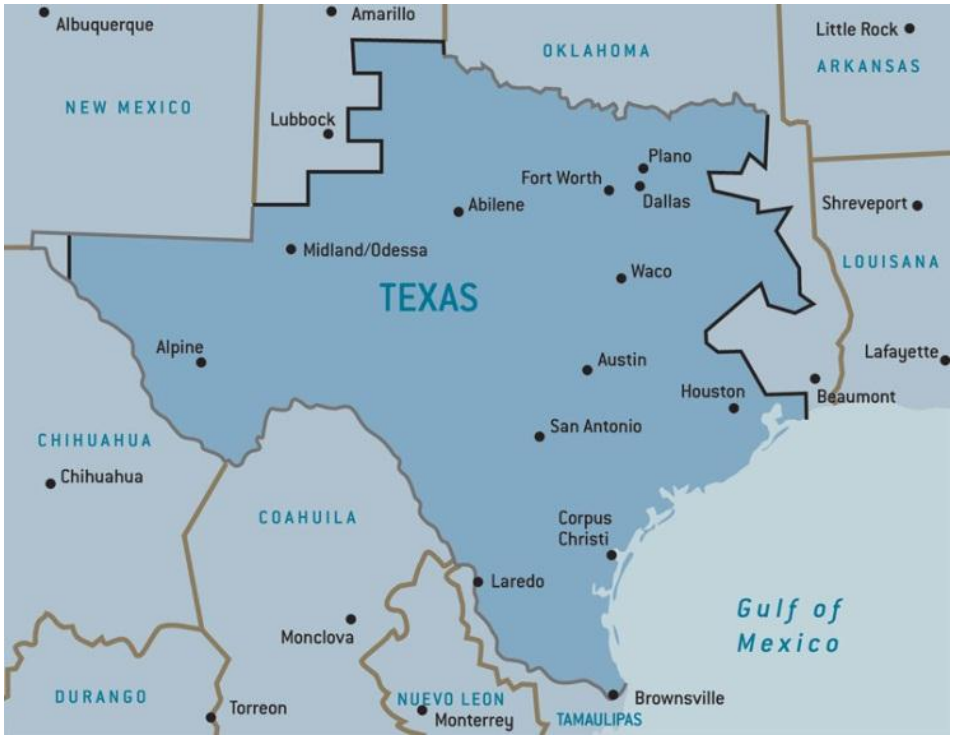
2.2.1. Creation and Development

Texas Interconnectivity System (TIS) has been set up to support the war effort during World War II. It provided excess power to the companies on the Gulf Coast, mainly engaged in the aluminum smelting. After the war, recognizing the benefits of the interconnectivity companies comprising TIS continued to cooperate under the commonly accepted guidelines and established two monitoring centers, one in the North and one in the South.

After the Northeast blackout of 1965, which left over 30 million people without electricity, the need to coordinate efforts of power utilities became critical. Electric Power Reliability Act of 1967 proposed the creation of a council which would coordinate the power market. Although not implemented the Act has lead Federal Power Commission (FPC) to propose the formation of a council comprised of regional coordination organizations. In response to the blackout and the FPC suggestions in 1968 the electric utility industry established the National Electricity Reliability Council (NERC). To comply with NERC requirements in 1970 TIS formed the Electric Reliability Council of Texas (ERCOT), and both organizations operated parallel until 1981 when TIS members transferred all operating functions to ERCOT and it became a central operating coordinator of the Texan electricity market. In order to deregulate the wholesale generation market, in the mid 1990's, Texas legislature amended the Public Utility Regulatory Act and substantially expanded ERCOT's responsibilities, these

included: promotion of the wholesale competition and facilitation of more efficient use of the power grid by all market participants.

Figure 3. Electric Reliability Council of Texas Region



Source: ERCOT

In 1996 ERCOT became the first electric utility industry Independent System Operator (ISO) in the US. Its goal was to promote nondiscriminatory transmission access, equitable interconnection process and customer protection. Three years later Texas Legislature passed Senate Bill 7 to deregulate electricity market and create retail competition.

LEGISLATIVE POLICY AND PURPOSE. (a) The legislature finds that the production and sale of electricity is not a monopoly warranting regulation of rates, operations, and services and that the public interest in competitive electric markets

requires that, except for transmission and distribution services and for the recovery of stranded costs, electric services and their prices should be determined by customer choices and the normal forces of competition. As a result, this chapter is enacted to protect the public interest during the transition to and in the establishment of a fully competitive electric power industry.

Senate Bill 7, Sec. 39.001 (a)

Goal of the retail competition and customer choice was to be achieved by the unbundling of electric utilities into generation, transmission/distribution and retail. Generation and retail sectors were granted freedom of competition, while transmission and distribution of electricity remained under the regulation of Public Utility Commission of Texas (PUCT).

UNBUNDLING. (a) On or before September 1, 2000, each electric utility shall separate from its regulated utility activities its customer energy services business activities that are otherwise also already widely available in the competitive market.

(b) Not later than January 1, 2002, each electric utility shall separate its business activities from one another into the following units:

(1) a power generation company; (2) a retail electric provider; and (3) a transmission and distribution utility.

(c) An electric utility may accomplish the separation required by Subsection (b) either through the creation of separate nonaffiliated companies or separate affiliated companies owned by a common holding company or through the sale of assets to a third party. An electric utility may create separate transmission and distribution utilities.

(d) Each electric utility shall unbundle under this section in a manner that provides for a separation of personnel, information flow, functions, and operations, consistent with Section 39.157(d).

Senate Bill 7, Sec. 39.051

Moreover, PUCT was given new functions, such as the ability to establish and enforce rules to protect customers from fraudulent, unfair, misleading deceptive or anticompetitive practices, oversee all providers of electric service and assess administrative and civil penalties for violations. ERCOT was given the authority to power scheduling, settlement, administration of a day-ahead ancillary services market, and administration of the retail customer-switching functions.

Table 8. Main Events in the Development of Texas Electricity Market

DATE	EVENT
1970	Electric Reliability Council of Texas created
1975	Public Utility Commission of Texas established
1992	Energy Policy Act
1995	Introduction of the Wholesale Competition
1996	ERCOT becomes Independent System Operator
1999	Senate Bill 7 passed
2001	ERCOT became a single control area
2002	Introduction of Retail Competition
2010	Introduction of Nodal Structure of the Electricity Market

Source: http://www.ercot.com/about/profile/history/#2010

The bill also eliminated state's integrated resource planning process and committed to the increase of renewable energy share in electricity generation, together with the establishment of a renewable energy credits trading program.

After passing Bill 7 in 2001, all control areas in the region were consolidated into one. Commercial functions were centralized with the intent to facilitate efficient market operations. Wholesale power sales between electric utilities became subject to new electric industry restructuring guidelines. Introduction of the competitive retail electricity market in 2002 allowed millions of consumers to choose their own providers. Since then Texas witnessed substantial volumes of consumer switching, with more than 2 million consumers switch their providers in less than five years.

Transition from a zonal framework of operation to a nodal structure of ERCOT was intended to provide price transparency, improve local price signals and centralize the electricity dispatch. Announced in 2006, the nodal system was to be setup by 2008, but due to the software implementation issues became operational only in 2010.

Today ERCOT operates the day-ahead and real-time markets. Day-ahead market is used in order to improve the procurement and delivery of electricity, as well as to ensure the reliability of the system. Here participating parties are able to optimize their bilateral contracts by scheduling both - ancillary and regular energy services. Qualified scheduling entities require less effort to find trading partners with load or generation and ancillary services ensure that ERCOT will have enough capacity to manage the grid next day. Information provided on a day-ahead market enables the reliability council to better forecast and manage congestion, more efficiently respond to *force majeure* situations. To meet the short-term load forecast and manage congestion between the nodes ERCOT uses balancing power procured at the real-time market.

Due to space limitations it is impossible to cover over forty years of history of the organization, therefore we refer the interested reader to explore a more comprehensive text - “Electricity Restructuring: The Texas Story” (Kiesling and Kleit, 2009) where one may find information on the preliminary treatment of the reform by the legislature, evolution of the wholesale market design, transmission policy, distributed generation, competitive performance of the wholesale market, retail restructuring and its design, and finally market monitoring by the Electricity Reliability Council of Texas.

2.2.2. Electricity Generation: Supply and Demand

The reliability and availability of the substantial generation capacity to satisfy demand for electricity is critical for the proper functioning of any economy on a local, state or country level. Households, hospitals, airports, continuous-process plants etc. heavily rely on uninterrupted supply of electricity. Following the requirements of PURA, ERCOT plans system-wide transmission, coordinates market transactions, ensures the reliability and adequacy of the regional electric network. It also ensures access to transmission and distribution systems for all buyers and sellers of electricity on nondiscriminatory terms. Along with transmission ERCOT coordinates efforts of private and municipal entities to provide additional generation facilities and their interconnection with the system. Deregulation of the market coupled with growing demand for electricity and relatively light regulatory environment attract power plant developers to Texas.

Table 9. Generation Capacity by Weather Zone and Ages of Plants in 2004

Weather Zone	Age of Plants (years)					
	> 50	49-40	39-30	29-20	19-10	<10
	MW	MW	MW	MW	MW	MW
Coast	161	2,100	7,159	3,876	3,852	5,234
East	0	210	2,533	4,158	305	2,748
Far West	2	256	545	0	580	1,910
North	106	257	1,054	40	998	2,244
North Central	468	2,326	4,214	2,106	4,324	5,642
South Central	362	655	3,159	2,591	1,280	4,613
Southern	247	558	1,268	1,533	0	2,981
West	84	268	1,226	276	724	327
Total	1,430	6,630	21,157	14,580	12,064	25,699

Source: ERCOT

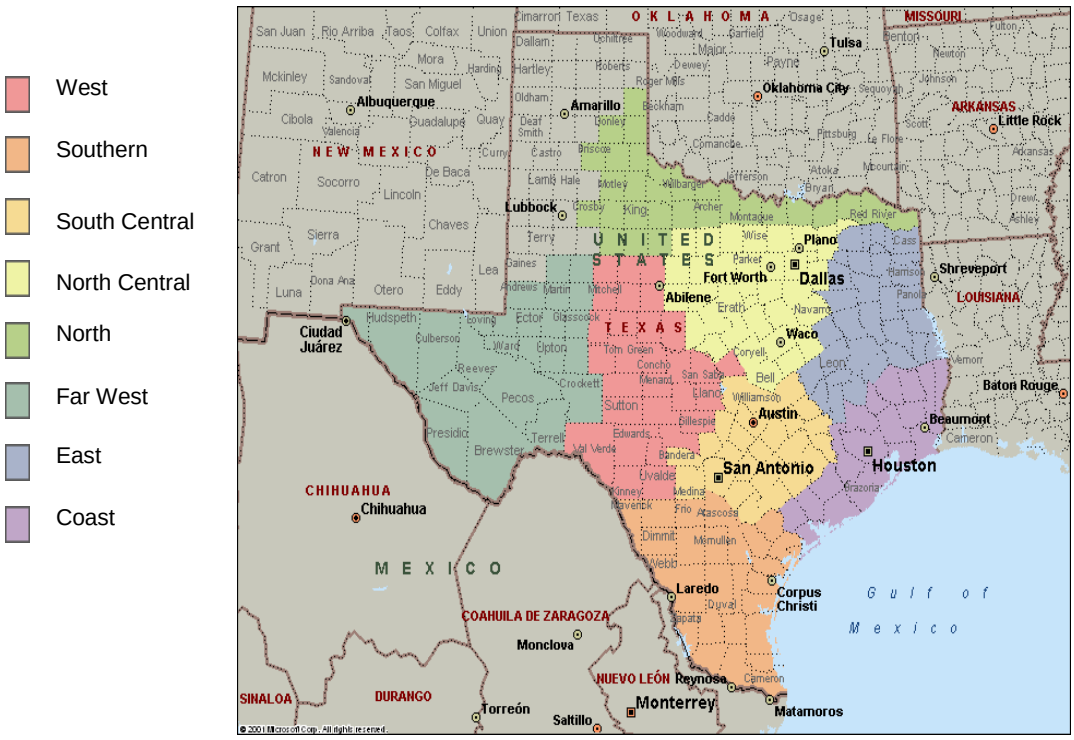
In compliance with PURA and PUCT Substantive Rules, ERCOT evaluates system constraints, studies transmission projects intended to relieve those constraints, recommends existing transmission projects that address some of the constraints and annually releases a report to inform the public on "existing and potential electric system constraints and needs"[10].

A 2004 review provided information on the age of infrastructure in different areas of Texas, known as weather zones (see Table 9 and Figure 4). We notice that more than half of the capacity was installed twenty or more years ago and only thirty per cent of generating capacity have been operating less than ten years. The majority of older equipment is located in the heavier populated areas such as Austin, Dallas-Fort Worth, Houston and San Antonio, accounting for over 60% of state's population. These are also

[10] Complete reports may be found at the ERCOT's webpage: http://www.ercot.com/news/presentations/

the areas where new equipment is being installed at a faster rate than elsewhere in the state.

Figure 4. ERCOT Weather Zones

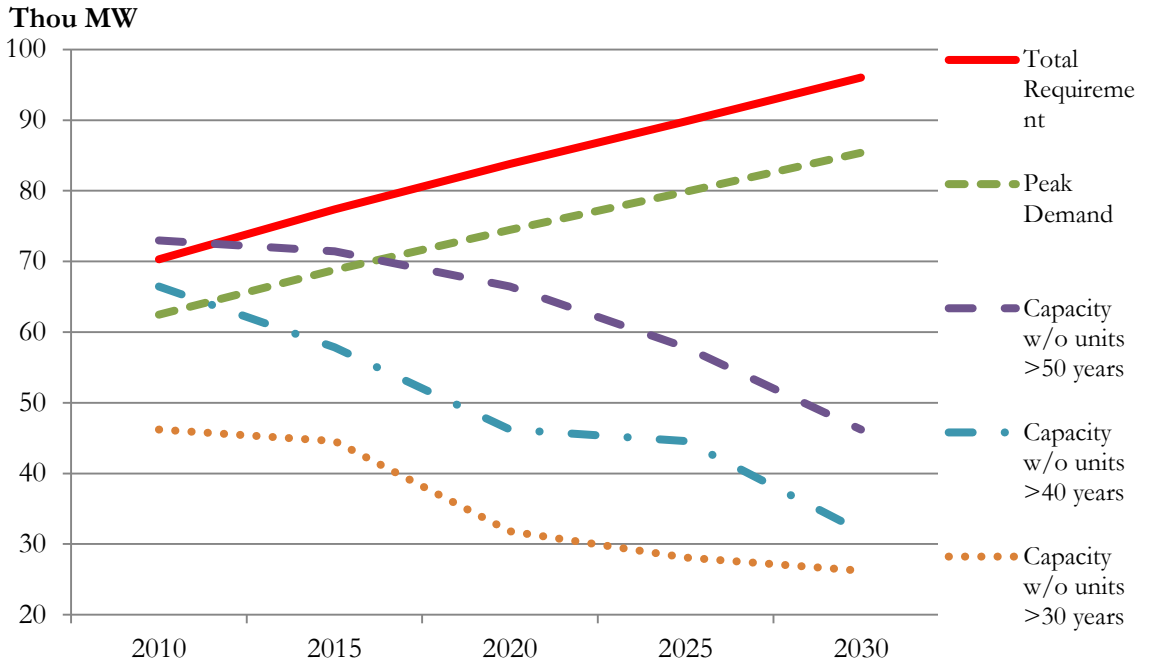


Source: ERCOT

To ensure the reliability of the system, as well as the availability of an adequate level of capacity to satisfy demand, ERCOT uses the so-called “reserve margin” in its estimations and forecasts. Reserve margin is defined as the percentage excess of the available generating capacity over peak demand in the system, or simply put – amount by which resources exceed maximum load. Originally target reserve margin was set to be 12.5% and existed until 2010, when it was raised by the Board of Directors to 13.75% for 2011 and years to come. By adjusting the methodology of reserve margin determination ERCOT recognized that generator’s contribution to reserve has to be evaluated based on

its availability rather than nameplate capacity. According to the President and CEO of ERCOT Trip Doggett, current reserve margin exceeds the target by approximately 3%, but with planned retirement of existing facilities will be declining, already not reaching the target by 3% in 2015[11].

Figure 5. ERCOT Generation Capacity and Demand Projections



Source: ERCOT

Figure 5, provides the forecast for expected peak demand and currently installed capacity with gradual retirement accounting for its age, calculated in megawatts (MW). We note that today's growing peak demand, given the forecast, exceeds the capabilities of the electricity generators whose facilities were installed less than forty years ago and other things being equal by 2017 is expected to overgrow the existing capacity. Total

[11] http://www.ercot.com/content/news/presentations/2011/DOGGETT%20-%20Austin%20Metropolitan%20Breakfast%20Club%2011-9-11.pdf

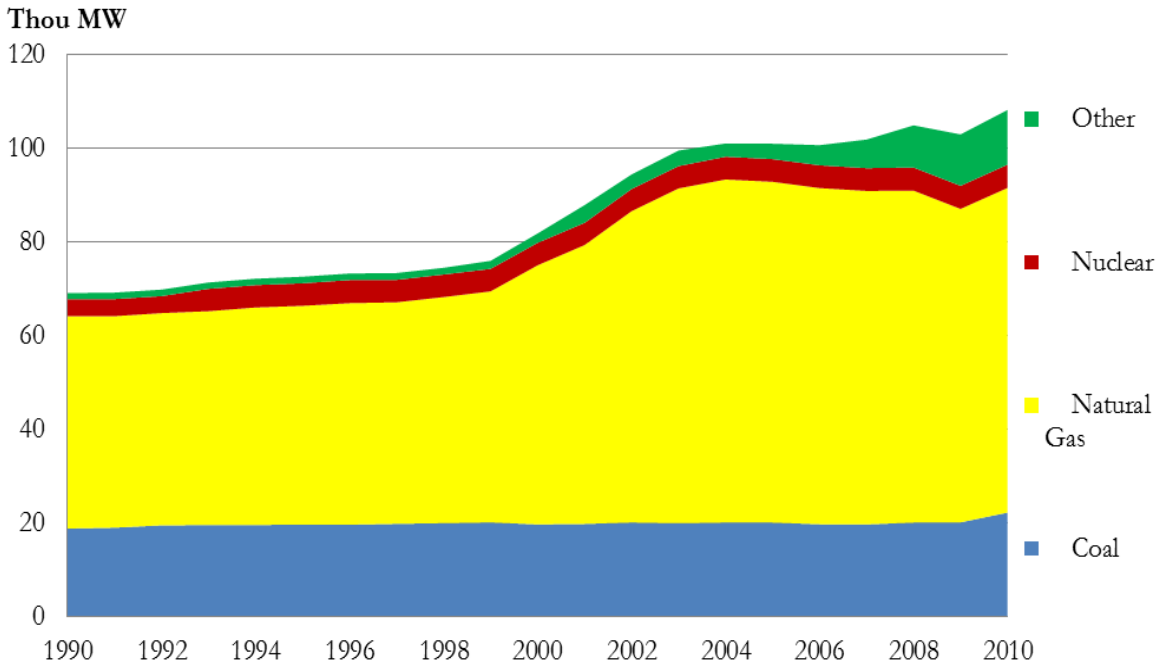
requirement (peak demand plus reserve margin) has already surpassed the capacity available in Texas.

Installation of new capital has had a changing dynamic in the growing market of Texas. While the investment and installation of new generators was slightly higher than the retirement and mothballing of older facilities in the 1990s, ERCOT experienced a sharp rise in capacity accumulation during 1999-2003 with more than 30% increase, driven mainly by the growth of facilities powered by natural gas (Figure 6). This phenomenon may be explained by Bill 7, which required market deregulation before it was passed. Electric utility companies expected changes in the market structure and had no incentives to expand their generating capabilities. In the second half of 1999 the majority of generating capacity was divested into the newly formed independent producers, with only several electric utilities belonging to municipalities were not given to the private sector. Electricity generation by utilities and independent producers followed the pattern presented by the installed capacity divestiture with the shift between the two occurring in the second half of 2001. The abundance of natural gas in Texas, as well as the ease of implementation and use of gas powered generating technology made it a fuel of choice by many entrants to the market. Due to higher competition between investors in the second half of 2000's, ERCOT is allocating more time to assess and approve best projects, which have led to a certain leveling off in the new installed capacity. One also notes that the amount of "other"[12] fuel generators has more than

[12] These mainly include renewables, such as: solar, wind, waste and biomass, geothermal and hydro power.

doubled and surpassed nuclear facilities in the total share of installed capacity. The slowest growth was observed in the coal generating capacity which increased in absolute terms by less than 18% over twenty year period, while at the same time growth of the industry constituted more than 56%.

Figure 6. Installed Generating Capacity, by Fuel

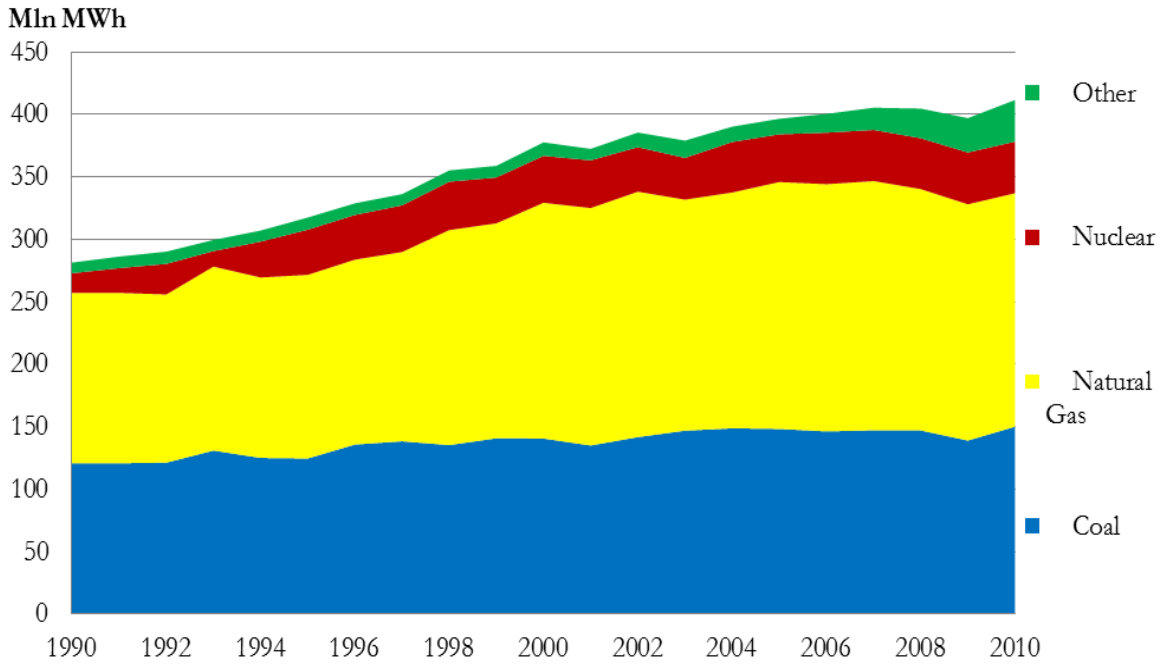


Source: EIA

Relative decrease in the share of coal powered generating capacity from 27.4% in 1990 to 20.6% in 2010 did not change the importance of coal in electricity generation, which today constitutes over 36% of total generation (Figure 7). Similarly, share of natural gas in generation has dropped by only 3% over twenty years to 45.4%, with its peak share of 51% in 2001, its capacity share was reduced by only 1.5%. Nuclear power generation remained almost unchanged and provides 10% of total electricity in the state.

Although a number of alternative energy projects were implemented in the second half of the 2000's and increased the share of alternative energy capacity, mainly wind, to over 10%, today these facilities produce only 8% of total electricity. Overall, electricity generation grew on average by 2% per year with 46% total increase between 1990 and 2010.

Figure 7. Total Electricity Generation, by Fuel

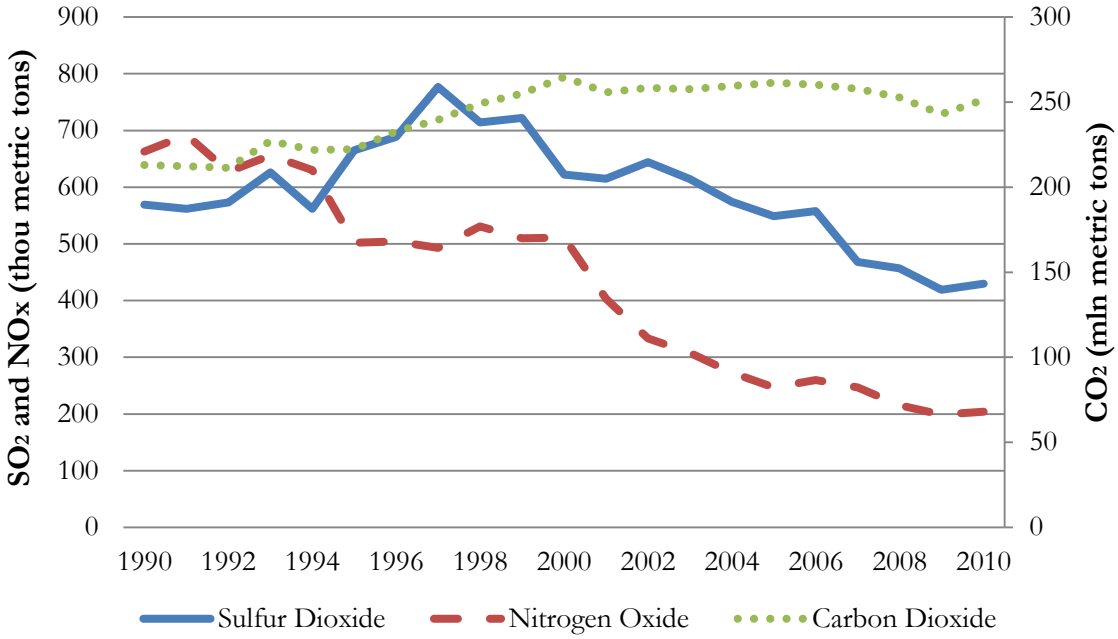


Source: EIA

According to the U.S. Energy Information Agency in 2010 Texas ranked first in the nation as carbon dioxide and nitrogen oxide emitter, and second in emission of sulfur dioxide. Although in relative terms – pounds of pollutant per megawatt hour (MWh) produced, Texas is ranked 22nd, 32nd and 28th respectively. Coal used by Texas power generation plants is mainly lignite, which has lower heat content as well as lower share of carbon as compared to other types of coal, at the same time sulfur content in some types

of lignite may be three times higher than in other types. Coal is responsible for the majority of pollution in the state: nearly 97% of SO_2, 52% of NO_x and 62% of CO_2 were generated at the coal fired power plants. Natural gas follows with 41% of NO_x and 38% of CO_2, while other fuels, such as landfill gas, sludge and solid waste, biomass and other agricultural byproducts generate 2.3% of SO_2 and nearly 7% of NO_x. The reduction of NO_x has seen a dramatic decrease by nearly 70% (Figure 8) and the SO_2 emissions have been lowered to 55% from their peak in 1997, CO_2 discharge has risen by 18% averaging nearly 0.8% growth per year during the period of observation.

Figure 8. Emissions

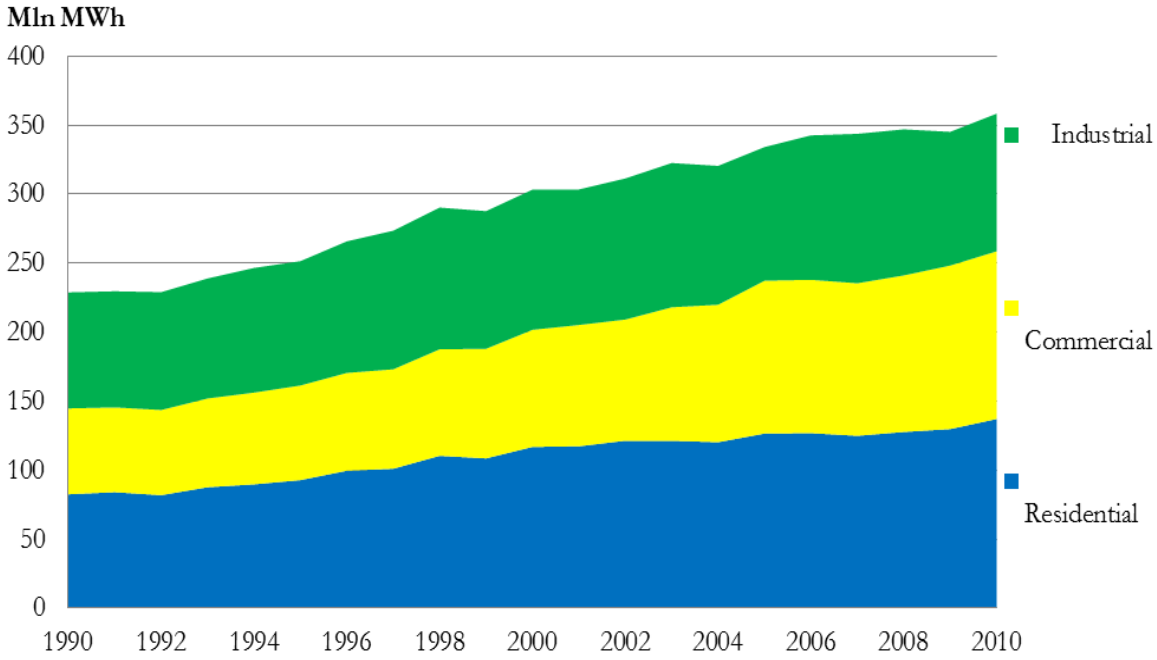


Source: EIA

The increased reliance on natural gas as well as stricter environmental rules have led to the reduction of SO_2 and NO_x emissions, polluting power plants were shut down and emission reducing equipment is becoming widespread. The increase in electricity

production and a sizeable reduction in acid rain pollutants together with some increase of carbon dioxide indicate an overall success in the implementation of new technology and government policy.

Figure 9. Retail Sales, by Consumer



Source: EIA

Excluding direct use and transmission losses all electricity is sold to the retail customers (Figure 9). The growth rate of actual electricity consumption is comparable to that of generation and constitutes 51% increase over twenty year period, reaching almost 360 mln MWh sold in 2010. Commercial electricity consumption nearly doubled and increased its share in total consumption by 7.7% to 33.9%, surpassing that of industrial consumption. The use of electricity in the industry increased by 18% and in 2010 constituted 27.8% of total energy consumption, at the same time manufacturing output expanded by almost 75% which may indicate the improvement of electricity saving

technologies in the sector. Residential use of electricity grew faster than population (50%) and increased by 66% compared to the 1990, while in relative terms it went up by only 3.5% in the total share of power consumption and constitutes the largest share of electricity consumption in the state.

2.3. Methodology

We closely follow the approach introduced in Sub-chapter 1.3 of this thesis. At the preliminary stage of analysis electricity generating plants are matched and separated into several groups based on their observable characteristics. Further, productive efficiency of each unit is estimated and aggregate group scores are constructed for the purpose of comparison. Malmquist productivity index is constructed and decomposed into efficiency and technology changes to evaluate the evolution of these components in the power generating market during the reform. Finally, influence of environmental factors on efficiency of electricity generators in Texas is estimated via truncated regression with bootstrap.

2.3.1. Propensity Score Matching

The reason for using Propensity Score Matching (PSM) in this study is twofold: first, for tackling the potential problem of omitted variable bias, when the explanatory variables in the regression setting (Step 2 of our approach) are not accounted for in the efficiency estimation stage (Step 1); and second, by the way of matching the decision

making units (DMUs) and reducing dimensionality of the problem whereby only units with similar characteristics are compared.

Propensity score method has been developed as a part of the literature dealing with treatment effects (Rosenbaum and Rubin, 1983). Usually, a binary treatment (may be extended to the case of multiple treatments) T is received by the DMUs, with units receiving treatment belonging to one group ($T = 1$) and units without treatment ($T = 0$) – to the other group (also known as "control treatment"). The response to different treatments denoted Y_1 and Y_0 respectively, coupled with observable variables Z (also known as "pre-treatment" variables) for each unit and their treatment are used in the evaluation of the so-called propensity score, which is the probability of treatment given all observable characteristics, $P(Z_i) = P(T_i|Z_i)$. The generalized model assumes observed treatment of a DMU T_i^* to be dependent on its characteristics, $T_i^* = f(Z_i'\beta) + \varepsilon_i$ and binary treatment:

$$\begin{cases} T_i = 1, & T_i^* > 0 \\ T_i = 0, & T_i^* \leq 0 \end{cases}$$

In the applied literature propensity score is estimated via logit or probit models. Under the logit setup ε is assumed to have logistic distribution and the score is estimated as:

$$P(Z_i) = E(T_i|Z_i) = \frac{1}{1 + e^{\lambda f(Z_i)}}$$

On the other hand, when one considers probit, we assume $\varepsilon \sim N(0, \sigma^2)$ and:

$$P(Z_i) = E(T_i|Z_i) = g(\lambda f(Z_i))$$

Next step of this procedure is to find DMUs similar in their characteristics summarized by the propensity score. Note that, since propensity score is a continuous measure between zero and one, chances of observing the same score for two DMUs are very low. In order to compare propensity scores a number of the so-called "matching" methods have been developed (e.g. Heckman et al., 1997), with the most popular being: nearest neighbor, radius, local linear and kernel matching. The main idea behind matching methods is to compare the scores of treated and control DMUs based on the "vicinity" of each unit, this approach may be utilized in finding single or multiple matches.

Matching procedures require the availability of treated and control groups, in the efficiency estimation we don't usually deal with such a distinction. Nevertheless, there exist several issues which are hard to tackle in the framework of linear programing which is often used in efficiency estimation. One such problem is the existence of outliers. Linear programming techniques arguably provide misleading results in the presence of outliers. Despite the ambiguity of labeling a particular observation an outlier several methods were proposed to deal with this problem. Andrews and Pregibon (1978) proposed a jackknife-like procedure - they calculated ratios of volumes spanned by the sample and its subsamples. Wilson (1993) extended this approach to multiple outputs, but was criticized by Simar (2003) for not being direction-specific and being based on influence function arguments as well as relying on the convexity assumption. Cazals et al. (2002) suggested a nonparametric estimator robust to outliers. All the above

mentioned approaches are based only on the input-output values observed for each DMU. In case of matching, one is able to take into account all the observed features of the DMU and thus be more precise in outlier detection. A second problem with the linear programming techniques is the use of indicator variables, which in our model may be easily introduced as a treatment. Moreover, the matching approach suggested here may accommodate multiple dummy-treatment variables, something none of the existing dummy variable-DEA models is capable of.

2.3.2. Data Envelopment Analysis and Malmquist Productivity Index

In our model each generator uses N inputs $x^j = (x_1^j, \dots, x_N^j)$ to produce M outputs $y^j = (y_1^j, \dots, y_M^j)$. In order to produce comparable results and despite the fact that different types of generators use different technologies (e.g. gas turbine vs. steam generator) we assume that technology used to convert inputs, into outputs, is accessible to all DMUs and can be characterized by the technology set T defined as:

$$T \equiv \{(x, y) \in \Re^N \times \Re^M \mid x \; can \; produce \; y\} \tag{17}$$

We also assume that technology characterization satisfies standard regularity conditions of the production theory which were provided in previous chapter. The production model is set up as output oriented presuming the maximization of generated electricity given fuel, labor and capital available for each generator. To find the inefficiency of a DMU we use the Farrell output measure of efficiency operating at (x, y):

$$\delta(x, y) \equiv \max_{\theta}\{\theta: (x, \theta y) \in T\} \quad (18)$$

DMUs are considered to be efficient when $\delta(x, y) = 1$ and inefficient if $\delta(x, y) > 1$. A measure $\left[1 - \frac{1}{\delta(x,y)}\right] \times 100\%$ represents a share of inefficiency.

The intertemporal measure of the Malmquist productivity index (MPI) measuring Total Factor Productivity (TFP) change between two time periods will be used for the evaluation purpose of market evolution taking place over time. Following Caves et al. (1982), the output-oriented Malmquist TFP change index between base period t and time $t + 1$ is given by:

$$M_o(x_t, y_t, x_{t+1}, y_{t+1}) = \left[\frac{\delta^t(x_{t+1}, y_{t+1})}{\delta^t(x_t, y_t)} \times \frac{\delta^{t+1}(x_{t+1}, y_{t+1})}{\delta^{t+1}(x_t, y_t)}\right]^{\frac{1}{2}}$$

here $\delta^t(y_{t+1}, x_{t+1})$ represents the distance of maximum proportional change in output required to bring observation (y_{t+1}, x_{t+1}) to the period t technology. Moreover, similarly to the parametric approach of Nishimizu and Page (1982), Färe et al. (1992) used a non-parametric method to decompose the MPI into changes in efficiency and technology:

$$M_o(x_t, y_t, x_{t+1}, y_{t+1}) = \frac{\delta^{t+1}(x_{t+1}, y_{t+1})}{\delta^t(x_t, y_t)} \times \left[\frac{\delta^t(x_{t+1}, y_{t+1})}{\delta^{t+1}(x_{t+1}, y_{t+1})} \times \frac{\delta^t(x_t, y_t)}{\delta^{t+1}(x_t, y_t)}\right]^{\frac{1}{2}}$$

we may write it as follows:

$$\Delta TFP = \Delta Eff \times \Delta Tech$$

That is, change in TFP is equal to the change in efficiency defined as a ratio of the Farrell technical efficiency in time t to the Farrell technical efficiency in time $t+1$ and change in technology – a geometric mean of the shift in technology between the two periods. Values greater than one indicate improvement or growth, while values smaller than one show the decline in productivity, efficiency and technology components.

Since true technology and therefore potential output is unknown one needs to estimate inefficiency measures of individual DMUs and subsequently evaluate changes in TFP based on the observed data. In this paper one of the nonparametric estimators is employed, namely - DEA. The convex hull of the technology set T obtained using DEA[13], for assumption of constant returns to scale (CRS) is defined as $\hat{T}$:

$$\hat{T} \equiv \{(x,y) \in \Re^N \times \Re^M | \sum_{k=1}^{n} z_k y^k \geq y, \sum_{k=1}^{n} z_k x^k \leq x,\ z_k \geq 0, k = \overline{1,n}$$

$\hat{T}$ is known to be a consistent estimator of a true technology under the assumption of CRS[14]. After estimating $\hat{T}$ one may easily obtain efficiency measures for each DMU $\hat{\delta}(x,y)$ using (18). Since the estimated technology set is a subset of the true technology, $\hat{\delta}(x,y)$ is downward biased, but known to be consistent and asymptotically unbiased estimator of $\delta(x,y)$.

[13] Originated by Farell (1957) and popularized by Charnes at al. (1978).
[14] Consistency of the DEA estimator is shown in Kneip at al. (1998).

2.3.3. Truncated Regression with Bootstrap

Having estimated individual efficiencies we proceed with the analysis of environmental factors which may be responsible for shifts in efficiency. Since scores obtained in the first stage only provide relative ranking of DMUs under consideration, understanding the influence of factors beyond control of DMUs is needed to evaluate the system under which they operate. Impact of external factors is evaluated using the regression model:

$$\delta_j = w_j\beta + \varepsilon_j, j = \overline{1,n} \tag{19}$$

where the inefficiency score δ_j is determined by the environmental factors w_j via the vector of parameters β and some statistical errors ε_j. Note, that the dependent variable in (19) is bounded by unity, so the distribution of the error term is restricted. Following Simar and Wilson (2007) we employ truncated regression with bootstrap to estimate the coefficients of the above regression. We assume that the error term is distributed as a truncated normal with zero mean and unknown variance and take the truncation point to be determined by $\varepsilon_j \geq 1 - w_j\beta$. The following model is estimated by maximization of the corresponding likelihood function, with respect to $(\beta, \sigma_\varepsilon^2)$:

$$\hat{\delta}_j \approx w_j\beta + \varepsilon_j, j = \overline{1,n} \tag{20}$$

where

$$\varepsilon_j \sim (0, \sigma_\varepsilon^2), \qquad \text{s.t. } \varepsilon_j \geq 1 - w_j\beta, \quad j = \overline{1,n} \tag{21}$$

Once the coefficients are estimated, parametric regression bootstrap is further used to obtain confidence intervals of the estimated parameters.

It has been pointed out (e.g. Wang and Schmidt, 2002) that when external variables influence inefficiency, the two step procedure is plagued by the omitted variable bias, which originates in the first step and then is carried over to the second. It was shown in the previous chapter that preliminary treatment of the dataset by matching DMUs on their propensity scores eliminates units which are not comparable on their "characteristics", these characteristics may as well be the external factors influencing efficiency. Furthermore, in Monte Carlo simulations the procedure described above produced results which are closely comparable to the truth. Our procedure, based on the results from Sub-chapter 1.3 will produce comparatively better results than other procedures currently available to the researcher.

2.4. Data and Models

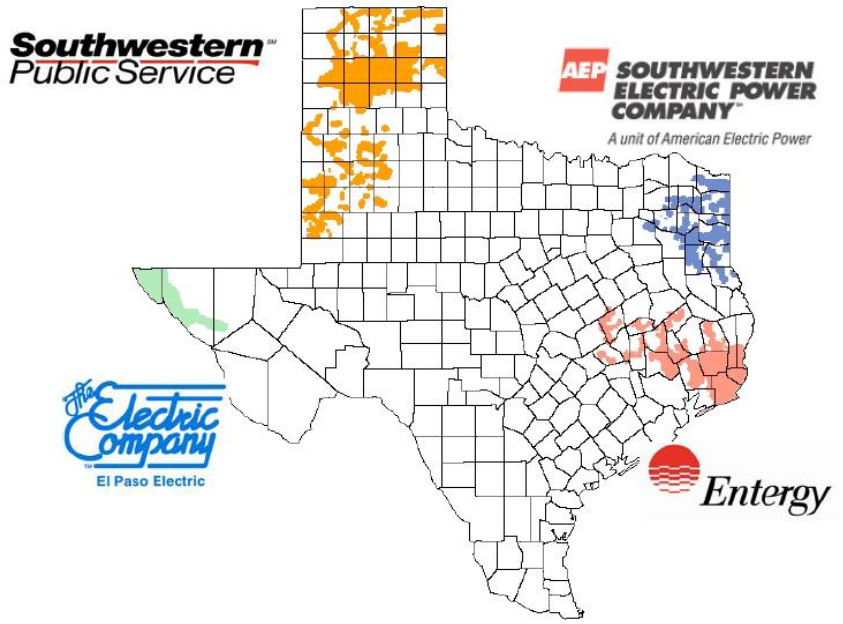
2.4.1 Data

The dataset used in this research was compiled based on the FERC Form 1, which became available on-line in the late 2000's[15]. Although U.S. Energy Information Administration provides a host of datasets which may be used in the analysis of

[15] FERC Form 1 data may be accessed at: http://www.ferc.gov/docs-filing/forms/form-1/data.asp

electricity markets, they lack uniformity of presentation as well as do not provide information on some of the variables that may interest us in efficiency analysis[16].

Figure 10. Electricity Generating Companies outside the ERCOT Grid



Source: Association of Electric Companies of Texas, Inc.

Our sample contains information on annual measures of output, three inputs (capital, labor and fuel) and five environmental variables at the generator level of each power plant. Three groups of generators may be identified in the data: vertically-integrated investor-owned electric utilities (e.g. American Electric Power Company, Centerpoint, Texas Utilities) in ERCOT, potentially subject to deregulation; vertically-integrated utilities outside the competitive Texas wholesale market (El Paso Electric

[16] EIA data may be accessed at: http://205.254.135.7/electricity/data.cfm

Company, Entergy-Gulf States Utilities, Southwestern Electric Power Company and Southwestern Public Service Company) which will be used as the control group; and municipal utilities in Texas (e.g. Austin Energy and City Public Service of San Antonio) - a "quasi control group" of utilities that have been exposed to competition in the wholesale generation sector but have not faced retail competition. One should note that municipal utilities have benefits over other generators in terms of exemption from some local taxes as well as weaker regulations that may be crucial in electricity generation market.

Regulated electric utilities were required to submit Form 1 to FERC, but since generation facilities were privatized, less information was provided and our sample shrank drastically for the year 2002 and remained small afterwards. Our dataset may be conveniently separated into three balanced subsamples: 1) within and outside ERCOT utilities; 2) municipal, within and outside ERCOT utilities; 3) municipal and outside ERCOT utilities. Each utility in the sample owns several power plants and since with time ERCOT companies divest them, balancing the data required omission of power plants not available for the time period under consideration. Eight years of data (1994-2001) is available for the first subsample, which contains information on 43 power plants within and 11 plants outside ERCOT's jurisdiction. 48 plants within ERCOT, 12 outside and 13 Texan municipal power plants over four years (1998-2001) comprise second subsample. And finally, third subsample consists of 12 outside and 11 municipal power plants observed during six years (1998-2003).

Plants in our sample represent over 50% of capacity installed in ERCOT and generate more than 50% of electricity. Summary statistics for three subsamples are presented in Table 10.

Table 10. Power Plant Summary Statistics

	Mean	Median	St. Dev.	Min	Max	Obs
Subsample 1						
Output (GWh)	3097.31	2010.62	188.52	0.72	20151.20	432
Capacity (MW)	770.95	575.00	34.91	40.00	3952.80	432
Labor (employees)	76.19	43.00	4.46	2.00	575.00	432
Age (years)	25.58	25.00	0.39	0.00	50.00	432
Peak Load (MW)	730.90	564.50	33.80	29.00	3879.00	432
Connection to Load (hours)	6802.94	7442.50	100.87	97.00	8784.00	432
Total Cost (mln $)	283.98	109.43	22.80	3.61	2675.45	432
Fuel Expenses (mln $)	87.08	60.46	4.51	0.10	568.30	432
Total Expenses (mln $)	100.53	67.27	5.20	0.25	630.89	432
Subsample 2						
Output (GWh)	2830.50	1690.97	217.39	2.64	20151.20	292
Capacity (MW)	729.50	563.50	40.78	27.50	3969.12	292
Labor (employees)	65.88	38.00	4.64	2.00	439.00	292
Age (years)	27.77	27.00	0.57	0.00	50.00	292
Peak Load (MW)	684.99	536.00	39.20	22.50	3879.00	292
Connection to Load (hours)	6576.32	7354.50	136.24	194.00	8784.00	292
Total Cost (mln $)	267.71	106.06	25.08	3.61	2414.12	292
Fuel Expenses (mln $)	83.11	49.65	5.31	0.25	513.87	292
Total Expenses (mln $)	95.60	62.24	6.14	0.98	595.21	292
Subsample 3						
Output (GWh)	2589.64	2028.42	248.40	1.69	12309.62	138
Capacity (MW)	663.67	544.00	44.86	27.50	2194.00	138
Labor (employees)	71.71	57.00	4.92	11.00	310.00	138
Age (years)	26.46	27.00	0.79	6.00	49.00	138
Peak Load (MW)	619.33	526.00	39.77	22.50	1885.00	138
Connection to Load (hours)	6589.42	7491.50	197.86	159.00	8784.00	138
Total Cost (mln $)	277.20	112.14	26.46	8.10	1274.72	138
Fuel Expenses (mln $)	66.73	49.62	4.96	0.14	303.48	138
Total Expenses (mln $)	78.35	61.91	5.50	1.41	325.59	138

The average installed capacity in all subsamples is between 660 and 770 MW. Power plants net generation - total generation excluding own use, is 2.5-3 thousand GWh each year. Number of people working at a power plant varies between 2 and 575, with an average of approximately 70 employees. The installed capacity at plants considered in

our work is on average 25 years old, with some units installed 50 years ago. The majority of power generators were connected to load for more than nine month per year, with some peak-load generators connected for less than a day and base-load generators connected the whole year. To build a power plant required on average 276 million and took nearly 91.5 million of 2010 dollars to operate. Expenditures on fuel comprised almost 87% of the total generation budget, with the remaining part devoted mainly to maintenance.

2.4.2 Models

In contrast to the engineering view of fixed proportions of capital and labor at a generating facility, a number of studies, e.g. Christensen and Green (1976) and more recently Rungsuriyawiboon and Coelli (2006) show that from the economic perspective electricity generators exhibit constant returns to scale (CRS). Following this observation the CRS production function with output orientation is set up, treating the objective of the power plant to generate as much electricity as possible given the resources. We hypothesize that after the announcement of the reform and introduction of the wholesale market in 1995 generators expecting the "unbundling" continued operating under same objectives, but slowed down the upgrades as well as the installation of new capacity and utilized all the existing capital, therefore improving efficiency. In the calculation of MPI and its decomposition changes of productivity are expected to be mainly driven by the increase of productive efficiency. In each available subsample power plants from different subgroups are matched based on the characteristics indirectly involved in the production process (e.g. time connected to load, current expenses, etc., see Table 10) and

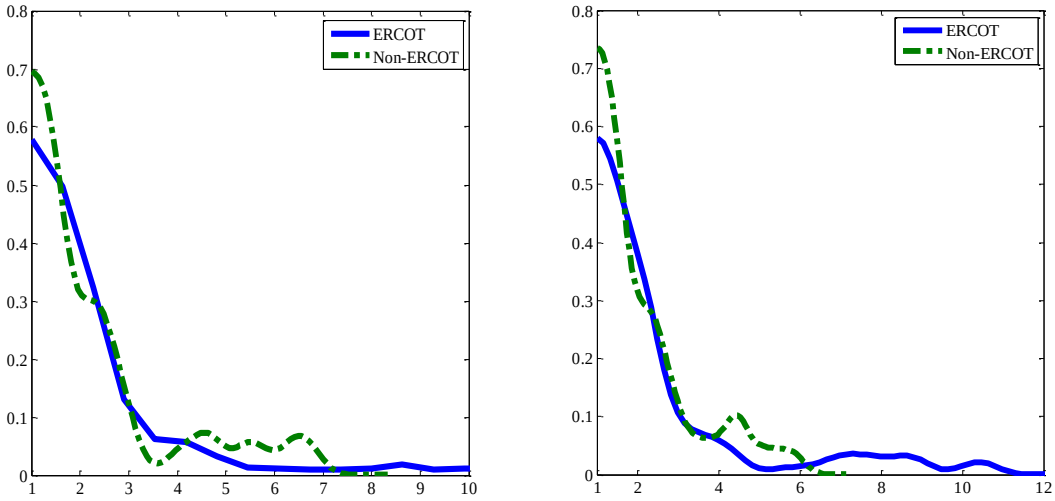
proceed with the comparison of similar units. Truncated regressions are estimated to access the impact of external factors on the productive efficiency of generating units.

2.5. Results

2.5.1. Vertically-Integrated Electricity Utilities in Texas

Comparison of utilities within the jurisdiction of different system operators may shed some light on the deregulation process of the Texan electricity market. Figure 11 shows that units both in and outside ERCOT operate on a similar level of efficiency, although when only comparing generators based on their observable characteristics - concentration near the production frontier increases. This observation may be explained by the fact that ERCOT sample contains more peak-load generators, i.e. power plants used only during peak demand for electricity, which are less efficient than base-load capacity and by excluding some of them efficiency in the sample improves. Aggregating individual efficiency scores into groups of relevant generators, as suggested by Simar and Zelenyuk (2007), we observe that overall ERCOT utilities performed worse than their counterparts in the sample: in 1994 all plants were similar in their production patterns and with time improved their efficiency. All ERCOT plants improved slower, but in a matched sample sharply reduced inefficiency in 1996 and subsequently grew more inefficient (Figure 12). Non-ERCOT generators operating in Texas are characterized by lower inefficiency and after the introduction of the wholesale market were shortly dominated by similar regulated generators.

Figure 11. Distribution of Efficiency Scores among Private Electric Generators



a) Unmatched b) Matched

To analyze the year-to-year developments of the market in terms of its efficiency and technology improvement Malmquist productivity index with its decomposition is calculated. Furthermore we use the bootstrap procedure to obtain confidence intervals for each of the components under consideration[17]; these results are reported in Table 4, with confidence interval provided in brackets.

Efficiency steadily fell between 1995 and 2001, with a burst in 1997 increasing by nearly 28%. It insignificantly changed during 1995-1997 and abruptly dropped before the opening of the retail competition. Bootstrap confidence intervals suggest that only changes observed during 1995-1997 are not statistically different from zero.

[17] 90% confidence intervals are obtained using 1000 bootstrap replications.

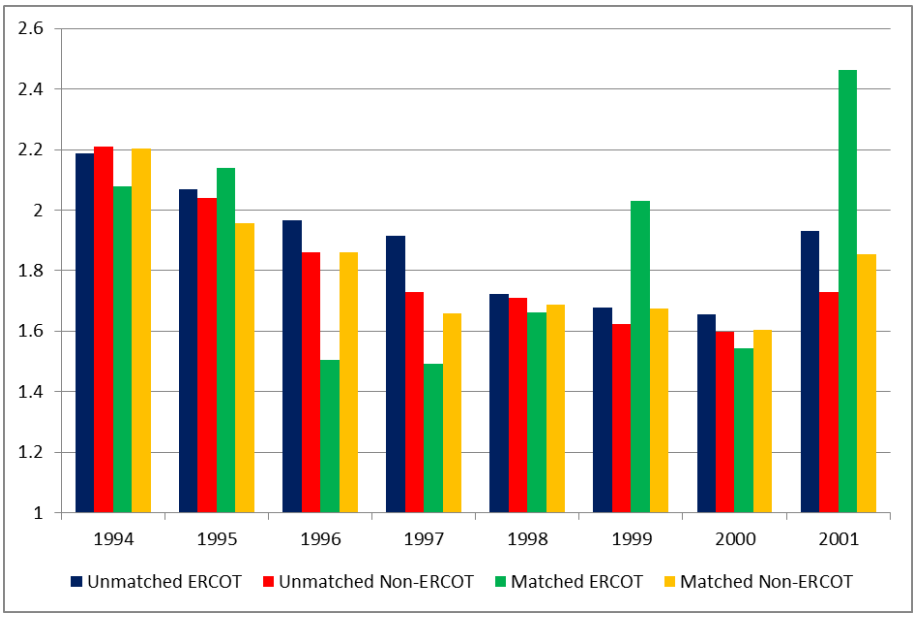
Table 11. Cumulative Indices of Efficiency, Technical and TFP Change. Regulated Generators

Year	Efficiency Change	Technical Change	TFP Change
94/95	1.051 [1.005; 1.097]	1.065 [1.054; 1.076]	1.119 [1.059; 1.180]
95/96	0.976 [0.929; 1.024]	1.020 [1.005; 1.036]	0.995 [0.936; 1.053]
96/97	0.989 [0.951; 1.028]	1.014 [0.999; 1.028]	1.000 [0.956; 1.045]
97/98	1.282 [1.005; 1.559]	1.045 [1.031; 1.060]	1.325 [1.005; 1.645]
98/99	0.956 [0.932; 0.979]	1.034 [1.027; 1.040]	0.987 [0.959; 1.015]
99/00	0.960 [0.932; 0.979]	0.986 [0.976; 0.996]	0.944 [0.901; 0.987]
00/01	0.837 [0.796; 0.878]	1.045 [1.021; 1.069]	0.863 [0.822; 0.904]

Technical changes, on the other hand, were slowly but steadily improving with an average of nearly 3% yearly growth with a small drop in 1999. All technology changes are significant at a 10% level. Combining changes in efficiency and technology TFP resembles the path of the efficiency changes: a big statistically significant jump in 1997 and the overall insignificant changes between 1995-1997 and 1998, as well as a substantial decline after 1999 describe the developments in TFP of Texas electricity generation market.

Next, let us consider the impact of external factors which are not directly involved in the production process, but have influence on power plant's operation. Besides the ability to compare generation units with similar characteristics, propensity score matching allows us to identify observations with similar external factors, which makes results from truncated regression more straightforward to interpret, since variables omitted from efficiency score calculation were implicitly accounted for during the matching procedure.

Figure 12. Aggregate Efficiencies of Regulated Generators in Different Markets



One of the assumptions about production process is the so-called "no free lunch" assumption, which states that when all production factors are zero then output produced is zero as well. Following this truncated regression has no intercept, meaning that when there is no production electricity generator is fully inefficient and its efficiency score is zero. Table 12 provides results from several specifications; the main difference between them is the inclusion of dummy variable controlling for ERCOT's jurisdiction and the consideration of fuel vs. total expenses incurred by the generating plant. As we mentioned earlier, total expenses variable is comprised of two main components – expenses on fuel and maintenance, therefore they are considered separately. With inefficiency as a dependent variable we observe that electricity generators in ERCOT are on average less efficient than their counterparts from other markets, but in both models the coefficient is not significantly different from zero.

Table 12. Truncated Regression Results for Regulated Generators

Variable	Model 1	Model 2	Model 3	Model 4
$ERCOT$	1.4154		1.5113	
Age	0.2792*	0.2652*	0.2708*	0.2691*
$\log(Capital)$	10.6475*	12.2108*	11.0737*	11.6421*
$\log(Demand)$	-10.5421*	-11.4704*	-10.5257*	-11.0886*
$\log(Employees)$	2.0397	2.8531*	3.0503*	3.1763*
$\log(Fuel)$	-3.0724*	-2.5220*		
$\log(Expences)$			-2.0988*	-2.4718*
$\log(Hours)$	-0.9468*	-1.1778*	-1.2061*	-1.1352*
$\log(TotCost)$	0.3222	-0.9638*	-1.1302*	-0.9942*

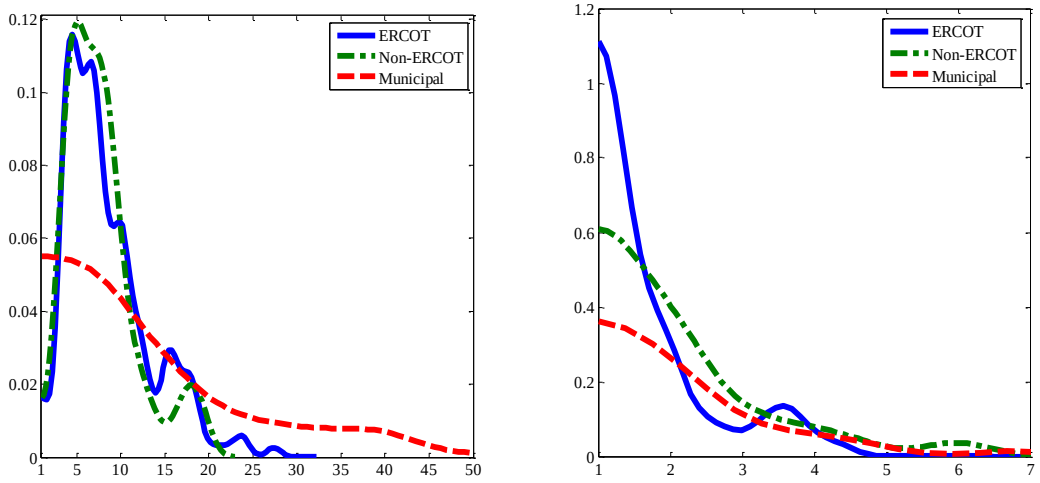
*-5%, **-10% significance

According to our models, age adds very little inefficiency to production, meaning that it is not a major factor for vertically integrated utilities. Additional capital, which may possibly be installed to satisfy peak load, adds to the inefficiency of the generator in Texas, due to a limited use. One also needs to remember that different units, depending on age and fuel may have different nameplate capacities and it would be best to control for this variable in a regression setting, unfortunately such data is not available. Peak demand and longer hours of connection to the grid, which if increased require generators to operate closer to their full capacity make power plants more efficient. Extra employees at the plant negatively influence generation efficiency of electricity, meaning that more workers possibly add to the problem of congestion. Fuel and maintenance expenses increase efficiency of power generators, as well as the total (fixed) cost of installed capacity and facilities which may indicate that more expensive (larger) plants designed to satisfy the base load are on average more efficient than the smaller ones.

2.5.2. Regulated and Municipal Electricity Utilities

With three groups at hand it is possible to not only to compare different regulated markets, but also to evaluate them against the unregulated environment. Highly outnumbered municipal generators provide their services to smaller communities in contrast with other generators in the sample, but are given preferences in taxation and lenient regulation. In the full subsample, generators are rather inefficient with several plants defining the frontier and the rest lagging substantially behind.

Figure 13. Distribution of Efficiency Scores among Private and Municipal Electric Generators



a) Unmatched b) Matched

Notably, facilities within ERCOT and outside its regulation have efficiencies very similar to each other and are widely spread, at the same time municipal generators are seen to be less efficient (Figure 13) and have facilities which are as inefficient as their counterparts. In the matched sample, as one would expect, efficiency distributions are characterized by similar patterns as before: larger concentration near the frontier and

smaller dispersion. One notices that among equal units ERCOT facilities are more efficient, while averages of comparable non-ERCOT and municipal generators are further away from the efficient production.

Limited by the time span decomposition of MPI provides information which supports previous findings (Table 13): technological change was small, positive and statistically significant during 1998-2001; efficiency dropped over the last years of the twentieth century. Finally, on a verge of a retail market opening efficiency and, consequently, productivity substantially fell, while technicaly improved by over 6%.

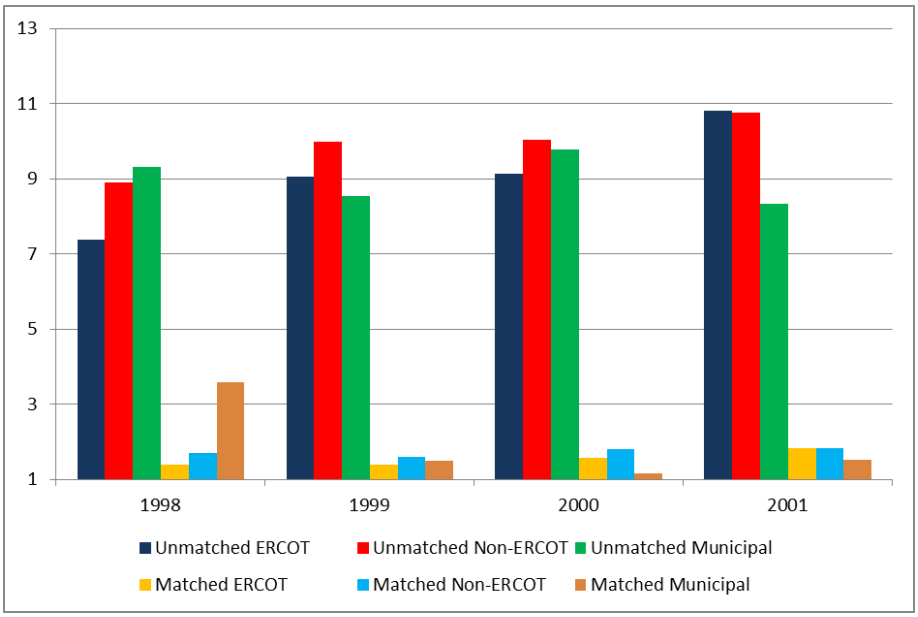
Table 13. Malmquist Index Decomposition. Regulated and Public Generators

Year	Efficiency Change	Technical Change	TFP Change
98/99	1.013 [0.903; 1.123]	1.037 [1.032; 1.041]	1.051 [0.935; 1.168]
99/00	0.979 [0.937; 1.020]	0.983 [0.975; 0.990]	0.958 [0.921; 0.995]
00/01	0.836 [0.798; 0.874]	1.063 [1.044; 1.083]	0.877 [0.845; 0.911]

Overall in 1998, municipal generators exhibited somewhat lower efficiency levels compared to their counterparts, but with time managed to improve and surpass ERCOT and non-ERCOT facilities, which themselves where very similar to each other.

A similar pattern is observed in the matched sub-sample, where considerably lower levels of inefficiency are presented by plants with comparable characteristics: municipal generators lag behind and in 1999 catch-up with the rest, at the same time efficiency of ERCOT facilities slightly deteriorates.

Figure 14. Aggregate Efficiencies of Regulated and Municipal Generators



Analyzing the impact of external factors we observe several similarities with previous subsample: additional installed capacity and age reduce productive efficiency of the plant, while peak load demand, expenses and the number of hours connected to load make plants operate more efficiently. ERCOT power plants produce electricity more efficiently when total expenses are taken into account, but when considering only fuel expenses - regulated Texan utilities are more inefficient. We note that in the specifications where total expenses are accounted for municipal generators also perform better, while in the model with fuel expenses municipalities have no significant advantage. Additional employees, possibly by creating congestion, are the detrimental factor to generation, while total cost of a plant usually helps to reduce inefficiency. Coefficients in models presented in Table 14 are statistically significant at 5% level if not indicated otherwise.

Table 14. Truncated Regression Results for Regulated and Municipal Generators

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$ERCOT$	0.0469*			-2.7444*		
$Municipal$		2.7377			-0.5267*	
Age	0.8949*	0.6051*	0.7902*	0.5174*	0.5370*	0.6774*
$\log(Capital)$	2.6709*	9.5799*	5.1788*	7.8042*	11.9444*	8.8864*
$\log(Demand)$	-0.0342*	-4.3001*	-0.4531*	0.0123*	-5.1862*	0.2448*
$\log(Employees)$	0.8008*	2.6804*	0.7694*	0.0318*	0.6947*	-0.0005*
$\log(Fuel)$	-4.0233*	-6.5059*	-5.1942*			
$\log(Expences)$				-11.933*	-10.324*	-10.510*
$\log(Hours)$	-4.9308*	-4.9414*	-5.5975*	-1.8513*	-2.5102*	-4.5665*
$\log(TotCost)$	0.9215**	-0.9845*	1.0994**	-3.7928*	-2.2255*	-2.4866*

*-5%, **-10% significance

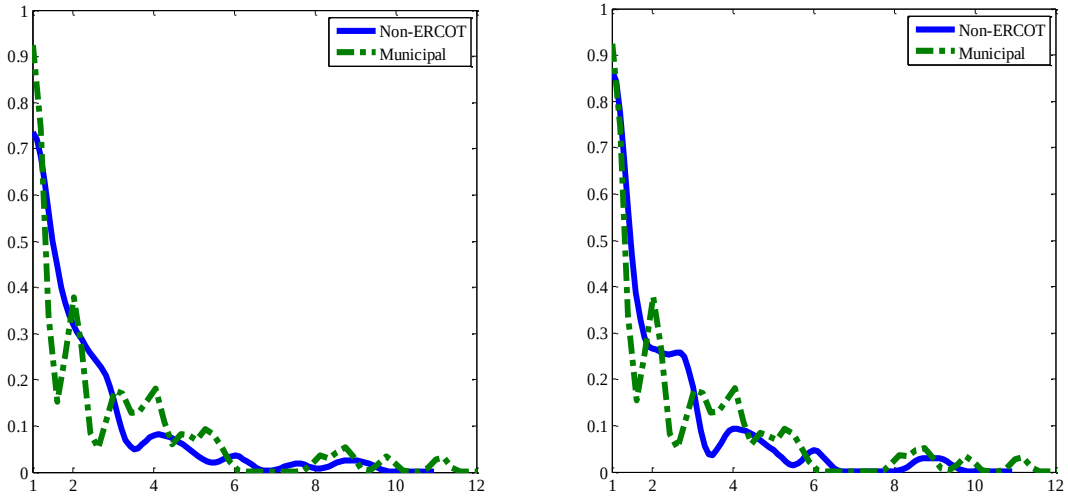
2.5.3. Non-ERCOT and Public Electricity Utilities

Finally we consider two groups that are not regulated by ERCOT – generators operating in the Southwest Power Pool, Western and Eastern Interconnections and Texan municipalities. One needs to note that nearly the same number of generators is available in two groups therefore we expect results for unmatched and matched subsamples be almost identical. Nonetheless we follow the procedures used in Sub-chapters 2.5.1 and 2.5.2.

In the unmatched and matched subsamples distributions are nearly identical and are characterized by the smaller dispersion if compared to the previous case. Peak load facilities locate themselves in the tails of distributions.

Decomposing the productivity index we observe further reduction of efficiency, technology and total factor productivity. Years 1998 and 2001 have seen some insignificant growth in all elements of productivity and deterioration of all indicators during other years in terms of both statistical and economic significance.

Figure 15. Distribution of Efficiency Scores among Non-ERCOT and Public Electric Generators



a) Unmatched b) Matched

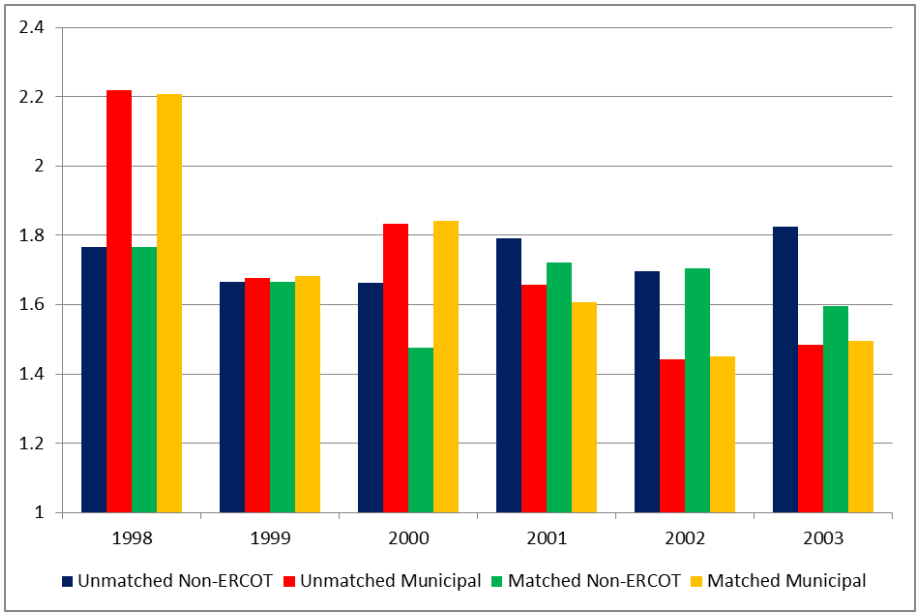
As anticipated, comparison of essentially same groups in the matched and unmatched subsamples provides similar aggregated figures: municipal electricity generators produce power more inefficiently and slowly catch-up with electricity generators outside ERCOT, especially in 1999 and 2001. At the same time regulated generators start with the reduction of inefficiency and continue its slow increase during 2001 and 2002 followed by a jump in 2003 for the unmatched case and see some reduction in the matched. These figures support previous observation about some reduction of productivity and its components.

Table 15. Malmquist Index Decomposition. Non-ERCOT and Public Generators

Year	Efficiency Change	Technical Change	TFP Change
98/99	1.199 [0.892; 1.506]	1.059 [1.043; 1.075]	1.278 [0.939; 1.616]
99/00	0.913 [0.867; 0.959]	0.999 [0.987; 1.010]	0.913 [0.865; 0.961]
00/01	0.888 [0.845; 0.931]	1.054 [1.035; 1.074]	0.934 [0.892; 0.974]
01/02	1.024 [0.929; 1.118]	1.053 [1.028; 1.077]	1.083 [0.968; 1.198]
02/03	0.947 [0.867; 1.027]	0.932 [0.927; 0.936]	0.882 [0.807; 0.958]

In the regression setting models try to control for the municipal generators as well as consider fuel and total expenses for electricity generation. Generators belonging to the municipalities on average have higher inefficiency. Greater age and more installed capacity add to the inefficiency as well. Interestingly, in the model where fuel expenses are considered, peak demand reduces inefficiency, while in the model with total cost coupled with the municipal dummy adds inefficiency, which may again be due to the congestion issues at the municipal level.

Figure 16. Aggregate Efficiencies of Non-ERCOT and Municipal Generators



More employees are shown to be detrimental to the inefficiency in the sample with fuel and maintenance expense, while when only fuel expenses are in play the relationship is negative. Since small size peak load electricity generating units comprise the sample, fuel and total expenses significantly reduce efficiency suggesting the possibility of congestion. Number of hours electricity power plants are connected to load

coupled with peak load demand make power plants generate electricity more effectively and total amount of money spent on the construction of the plant reduces productive inefficiency, suggesting that larger size municipal and non-ERCOT power plants operate closer to its potential. All coefficients have bootstrap confidence intervals supporting their significance, with the exception of several coefficients of installed capacity, peak load demand, employees and fuel expenses (Table 16).

Table 16. Truncated Regression Results for Non-ERCOT and Municipal Generators

Variable	Model 1	Model 2	Model 3	Model 4
$Municipal$	3.2790*		2.7139*	
Age	0.4683*	0.4990*	0.4349*	0.2577*
$\log(Capital)$	4.0519	5.5934*	-3.7416*	9.5238*
$\log(Demand)$	-2.2477	-1.7116	7.8209*	-4.7230*
$\log(Employees)$	1.7768	-1.2406*	0.2846*	4.0308*
$\log(Fuel)$	2.3763*	1.4539		
$\log(Expences)$			1.8807*	0.4702*
$\log(Hours)$	-2.4859*	-2.2698*	-2.6740*	-2.4632*
$\log(TotCost)$	-4.475*	-4.9285*	-5.6354*	-7.3158*

*-5%, **-10% significance

2.6. Conclusions

Electricity generating market in Texas has seen rapid growth in new generating capacity driven by consumer demand. Seen by the legislature as abusing market power vertically integrated utilities were to be unbundled and the newly created entities engaged in wholesale and retail competition. Rising electricity prices, attributed to the rise in natural gas prices by some, were negatively received by the public. This issue attracted a lot of attention among scientists and the demand side of the issue has been studied extensively. On the other side electricity generation was not studied in the academic

literature. Our intention is to fill this gap and at least partially analyse the supply side of the market.

Grouping electricity generators into vertically integrated utilities overseen by ERCOT and the ones operating under different regulators as well as municipal power plants we have found that although production of electric power has been rising and the inefficiency of all types of plants has been slowly deteriorating, after the year 2000 it reverted back to the growing path. Being able to match utilities from different regulatory regions we compare generating units based on similarities of their observable characteristics, such as installed capacity, number of workers, time connected to load, peak load demand etc. It was found that ERCOT facilities rarely outperform similar units from the non-ERCOT regulated group and in general operate on a similar efficiency level. Municipal generators exhibit higher aggregate inefficiency among comparable units with significant improvements after year 2000.

In contrast to our expectation, decomposition of the productivity index has shown rise in efficiency only in 1997 and its deterioration during other periods, meaning that rising competition did not significantly alter the utilization of resources available to power generators. Positive technological changes were the main significant driving force of the TFP over the years. Starting in 1998-1999 all facilities experience great negative efficiency change and as a result the reduction of total factor productivity.

The evaluation of the external factors, with previously matched subsamples, supports the idea that facilities with older and larger stock of capital tend to be more inefficient. This finding is plausible, since generators expect to be connected during the

peak hours when the demand exceeds the base load and therefore use extra capacity, which if idle adds to the inefficiency. Our model also supports the hypothesis that longer connection to load and higher peak demand enable generators to utilise their capacity better and thus produce more output, i.e. improving overall efficiency. All power plants seem to be losing efficiency by hiring additional workers, suggesting a possible congestion issues connected with this particular input. Expenses on fuel and maintenance reduce inefficiency, while in the sample with small generators (3^{rd} subsample) they slightly hurt productive efficiency. According to our models total cost of facilities at the power plant usually positively translates into efficiency, suggesting that larger facilities perform better and since the majority of plants in the dataset are large, base load generating utilities this result is plausible.

With the deregulation of the market came the reduced responsibility of private generators to report data to the authorities, which prohibited us from making a more thorough analysis – a comparison of private and vertically integrated utilities. Despite that there are several possible extensions to our work. First, analysis of other parts of the deregulated electricity market - i.e., transmitters and retailers, which will help us better understand the developments in Texan electricity market after the deregulation. Second, a larger scale study of comparison between ERCOT and other system operator jurisdictions will provide a better benchmark and therefore provide an assessment of performance of ERCOT and other regulated and deregulated markets.

Chapter 3

Semiparametric Estimations with Shape Constraints

3.1. Introduction

Economic theories often provide useful guidance on the modeling of real world data. For instance, utility function associated with rational preference is monotone. In addition, under convex preference, we obtain quasiconcavity. Demand functions of normal goods are downward sloping. Under the duality theorem, profit functions are concave in output price, while cost functions are monotonically increasing and concave in input price. Economists, when trying to model economic relationships, face two challenges, fidelity to economic theories and flexibility in functional forms. These two goals are often at odds; conformity to theorems usually dictates rigid functional forms, while flexible parameterizations sometimes lead to counter-intuitive predictions.

Aiming to address these two issues simultaneously, Wu and Sickles (2012) and Wu, Sickles, and Demchuk (2012) present a flexible semiparametric estimator that

incorporates shape constraints. They focus on the functional relationships with two constraints: monotonicity and concavity, since this class of functions occurs most frequently in economic studies. Functional relationships that possess either one of these two constraints are special cases of our estimator. Convexity can be easily accommodated by a simple negation of one parameter in our model. In what follows I detail the findings of Wu and Sickles (2012) and Wu, Sickles, and Demchuk (2012) and point to new directions and future work on this important issue in the econometric modeling of nonparametric relationships that appear in many economic settings.

Several approaches have been used to incorporate range/shape restrictions in estimations. A simple approach is transformation of variables. For instance, logarithmic transformation is commonly used to ensure positiveness of the predicted dependent variables; the Box-Cox transformation offers a more flexible alternative. Monotonicity can be achieved by special functional forms. For example, in the estimation of production functions, Cobb-Douglas, constant elasticity of substitution (CES), trans-log, generalized Leontief are popular choices due to their simplicity and some desirable properties.

To avoid rigid functional forms, semiparametric and nonpamametric methods have been modified to accommodate shape restrictions. An early example is Brunk's (1955) isotonic estimator, which essentially produces a monotone step function. Mukerjee (1988) and Mammen (1991) develop kernel-based isotonic regression techniques which consist of a kernel smoothing step and an isotonization step to ensure monotonicity. Instead of isotonization, Hall and Huang (2001) suggest a penalized kernel method to achieve monotonicity. Their method is further generalized by Racine and

Parmeter (2008) to allow for general constraints. An alternative to these kernel-based methods employs constrained smoothing splines. See e.g. among others, Ramsay (1988), Kelly and Rice (1990), and Mammen and Thomas-Agnam (1999). A third approach entails rearranging or sharpening the data or predictions (see e.g. Braun and Hall (2001) and Chernozhukov, et al. (2007)).

In this study we combine the transformation approach with flexible semiparametric estimations. Suppose $y = f(x)$ is a smooth monotone function of $x \in [0,1]$. Ramsey (1998) proposed an integral transformation to model monotone functions:

$$f(x) = \int_0^x \exp\big(g(s)\big)\, ds, \tag{22}$$

where g is a square integrable function on [0,1]. Since $f'(x) = \exp\big(g(x)\big) > 0$ and $f''(x) = f'(x)g'(x)$, it suggests that to impose both monotonicity and concavity constraints, we can augment (22) with a monotonicity constraint on $g(x)$. Consider

$$f(x) = \int_0^x \exp\Big(-\int_0^s g(t)dt\Big)\, ds. \tag{23}$$

We then have $f'(x) = \exp(-\int_0^x g(t)dt) > 0$ and $f''(x) = -f'(x)g(x)$, suggesting that $f''(\cdot) < 0$ if $g(\cdot) > 0$. Common candidates of g include $g(x) = x^2$ and $g(x) = exp(x)$. Other choices are certainly possible. In particular, we will show below that $g(x) = x^2$ is appealing for the proposed method due to theoretical and practical reasons.

The parameterization (23) can be characterized by the following differentiable function

$$g(x) = -\frac{f''(x)}{f'(\mathrm{x})},$$

whose general solution is given by

$$f(x) = \beta_0 + \beta_1 \int_0^x \exp\left(-\int_0^s g(t)dt\right) ds.$$

Consequently given an iid random sample $\{Y_i, X_i\}_{i=1}^n$, we consider the following parameterization for modelling a monotone and concave functional relationship

$$Y_i = f(X_i) + u_i = \beta_0 + \beta_1 \int_0^{X_i} exp\big(-\int_0^s g(t)dt\big)ds + u_i \quad (24)$$

where $g(\cdot) > 0$ and u_i is an error term with mean zero and finite variance σ^2. Furthermore, we will model $g(\cdot)$ with $g(h(\cdot))$, where h is a square integrable function defined on [0,1] free of constraints.

3.2. Estimator

One major advantage of the transformation-based approach to incorporate constraints is that we can transform a constrained problem into an unconstrained one. In our case, this boils down to the modeling of h. Lacking theoretical guidance we select to model h using flexible nonparametric estimators. More specifically, we use the spline estimator since it is relatively easy to implement additive structures in multiple regressions and to embed smoothers in nonlinear functionals with spline methods. Compared with the power series estimators, the spline estimators are piecewise polynomials and do not suffer from the

oscillations associated with power series.

Let $0 < k_1 < \cdots < k_M < 1$ be a series of knots of the spline basis functions. We consider the truncated polynomial splines in this study. For instance, the truncated power basis of degree p is given by

$$\Phi(x) = (1, x, \ldots, x^p, (x - k_1)_+^p, \ldots, (x - k_M)_+^p)^T,$$

where $(x)_+ = max(x, 0)$. Truncated polynomial splines are often transformed into B-splines that facilitate theoretical analysis and numerical evaluations. See, e.g., De Boor (2001) for the constructions and properties of B-splines.

Suppose Φ is a P-dimensional basis functions with $P = 1 + p + M$. Define $h(x) = c^T\Phi(x)$ with c being a P -dimensional vector of coefficients. We consider the following model

$$Y_i = f(X_i) + u_i = \beta_0 + \beta_1 \int_0^{X_i} \exp\left(-\int_0^s g(h(t))dt\right)ds + u_i \qquad (25)$$

We need the constant β_0 and 'slope' β_1 here for identification, since the paramerization of f does not allow for free location or scale parameters. To see this, consider the simplest case when $g(x) = a$, a non-zero constant. It follows that $f(x) = (1 - exp(-ax))/a$, whose location and scale can not independently vary.

Model (25) is a semiparametric model with two parametric coefficients and a nonparametric smoother f. To balance fidelity to the data (as measured by sum of squared residuals) and smoothness of the estimator, we adopt the approach of penalized

spline estimation. An alternative to the penalized spline method is the regression spline estimator, which balances the goodness-of-fit and smoothness trade-off through judicious selection of spline functions. The selection of basis functions for regression splines can be a daunting task, especially in multiple regressions.[18] In contrast, the penalized spline estimator uses a relatively generous spline basis, whose coefficients are penalized to improve smoothness. We choose the penalized spline approach because its penalty to a large degree is governed by a single smoothing parameter and therefore easier to implement.

We estimate model (25) by minimizing the sum of squared residuals plus a penalty on the roughness of f. The objective function is given by

$$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - f(X_i; c))^2 + \lambda D(f), \tag{26}$$

where $D(f) > 0$ reflects the roughness of f. A popular choice of the penalty is the integrated squared qth derivative of f, where $q \leq p$ for the pth degree truncated power series splines. In particular, $q = 2$ is commonly used, which leads to the natural cubic spline in smoothing splines.

For linear spline models, one can in principle specify the basis functions and the penalty term separately. For nonlinear models, careful choices of the parameterizations and penalty term can sometimes simplify the estimator considerably. In our case since

[18] Suppose we consider P possible basis functions. A complete subset selection entails 2^P evaluations of candidate models.

$g(x) = -f''(x)/f'(x) > 0$, a natural choice of the penalty is the integrated relative curvature; that is, $D(f) = \int_0^1 g(x)dx = -\int_0^1 f''(x)/f'(x)dx$. We notice that the penalty on the relative curvature penalizes not only the curvature of f but also small values of f'. Consequently, it prevents the 'boundary' solution where $f' = 0$.

3.3. Algorithm

Denote the solution to the proposed nonlinear estimation (26) by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ and $\hat{c}$. Let $m(x) = \int_0^x exp\big(-\int_0^s g\big(h(t)\big)dt\big)ds$. It follows that $D(f) = D(m)$. Define $\widehat{m}(X_i) = m(X_i; \hat{c})$ and $g'(x) = dg(x)/dx$. Replacing β with $\hat{\beta}$ and applying Taylor expansion to m in (26) with respect to c around $\hat{c}$ yields

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 m(X_i; \hat{c}) - \hat{\beta}_1 \hat{Z}_i(c - \hat{c})\right)^2 + \lambda D, \tag{27}$$

where

$$\hat{Z}_i = \frac{\partial \widehat{m}(X_i)}{\partial \hat{c}} = -\int_0^{X_i}\left\{\int_0^s \big(\Phi(t)g'\big(\hat{c}^T\Phi(t)\big)dt\big)\int_0^s exp\left(-g\big(\hat{c}^T\Phi(t)\big)\right)dt\right\}ds.$$

The first order condition of (27) with respect to c is then given by

$$-\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1\hat{Z}_i^T\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1\widehat{m}(X_i) - \hat{\beta}_1\hat{Z}_i(c - \hat{c})\right) + \lambda D' = 0, \tag{28}$$

where

$$D' = \frac{\partial D}{\partial c} = \int_0^1 \left(\Phi(x) g'\left(\hat{c}^T \Phi(x)\right) dx\right)$$

Next, denote $\widehat{D} = D(\hat{m})$ and $\widehat{D}'$ and $\widehat{D}''$ its first and second derivatives with respect to c evaluated at $\hat{c}$. Taking Taylor expansion of D' with respect to c around $\hat{c}$ yields

$$-\frac{1}{n}\sum_{i=1}^n \hat{\beta}_1 \hat{Z}_i^T \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{m}(X_i) - \hat{\beta}_1 \hat{Z}_i (c - \hat{c})\right) + \lambda \widehat{D}' + \lambda \widehat{D}''(c - \hat{c}) \approx 0. \quad (29)$$

Define $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{m}(X_i)$. Plugging $\hat{u}_i$ into (28) and rearranging terms yields

$$\left(\frac{1}{n}\sum_{i=1}^n \hat{\beta}_1^2 \hat{Z}_i^T \hat{Z}_i + \lambda \widehat{D}''\right)(c - \hat{c}) \approx \frac{1}{n}\sum_{i=1}^n \hat{\beta}_1 \hat{Z}_1^T \hat{u}_i - \lambda \widehat{D}'. \quad (30)$$

Expression (30) suggests a Gauss-Jordan algorithm to solve for the proposed estimator. Let $\hat{c}_-$ be the current estimate of c and $\hat{m}(X_i)$, $\hat{Z}_i$, $\widehat{D}'$, $\widehat{D}''$ and $\hat{u}_i$ be calculated with $c = \hat{c}_-$. We calculate $\hat{\beta}$ via the least squares by regressing Y on $\hat{m}(X)$. Denote $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_n)^T$ and $\hat{Z} = \left(\hat{Z}_1, \ldots, \hat{Z}_n\right)^T$. Holding $\hat{\beta}$ constant, we then update c according to the following formula:

$$c = \hat{c}_- + \left(\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \widehat{D}''\right)^{-1} \left(\frac{1}{n}\hat{\beta}_1 \hat{Z}^T \hat{u} - \lambda \widehat{D}'\right). \quad (31)$$

$\hat{\beta}$ and $\hat{c}$ are updated alternatively in this fashion until convergence. The global concavity of m ensures the existence and uniqueness of the solution.

Remark 1. The penalty terms $\widehat{D}'$ and $\widehat{D}''$ generally depend on the current estimate $\hat{c}_-$ and therefore need to be calculated anew at each stage of the updating. This updating process

is simplified considerably when $g(x) = \frac{1}{2}x^2$. Recall that $h(x) = c^T\Phi(x)$. Define $K = \int_0^1 \Phi(x)\Phi^T(x)\,dx$. It follows that $D(m) = \frac{1}{2}c^TKc$ and the updating formula (28) simplifies to

$$\hat{c} = \hat{c}_- + \left(\frac{1}{n}\hat{\beta}_1^2\hat{Z}^T\hat{Z} + \lambda K\right)^{-1}\left(\frac{1}{n}\hat{\beta}_1\hat{Z}^T\hat{u} - \lambda K\hat{c}_-\right).$$

Thus under quadratic g, the penalty matrix remains constant during the updating process. Moreover, the Taylor expansion given by (29) is exact.

3.4. Inferences

Despite the popularity of penalized spline methods, their theoretical properties have not been well understood. Some earlier results were provided in Wand (1999) and Aerts et al. (2002), who made the simplifying assumption that the dimension of the spline basis is fixed. Hall and Opsomer (2005) investigated this problem using a white noise representation. Claeskens et al. (2008) showed that dependent on the assumption that is formulated from the number of knots, the asymptotic properties of penalized splines are either similar to those of regression splines or to those of smoothing splines.[19] Kauerrmann et al. (2009) studied the asymptotic properties of penalized splines under the first scenario in the framework of generalized linear models. Li and Ruppert (2008) used

[19] Smoothing spline is a special case of the penalized spline estimator where the number of basis functions equals the number of observations. See, e.g. Wahba (1990) for general treatments of smoothing splines.

the device of equivalent kernel to study the second scenario.

Claekens et al. (2008) indicate that faster convergence rates are obtained under the scenario analogous to the regression splines. We focus our analysis on this framework. To facilitate the derivation, we first present an alternative representation of solution (31). Define $\tilde{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{m}(X_i) + \hat{\beta}_1 \hat{Z}_i \hat{c}_-$. Plugging $\tilde{Y}_i$ into (28) and rearranging terms yields

$$\left(\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \widehat{D}''\right) \hat{c} \approx \frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} + \lambda\left(\widehat{D}' - \widehat{D}'' \hat{c}_-\right),$$

where $\tilde{Y} = \left(\tilde{Y}_1, \ldots, \tilde{Y}_n\right)^T$. Holding $\hat{\beta}$ constant, we can update c using the following alternative formula:

$$\hat{c} = \left(\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda \widehat{D}''\right)^{-1} \left(\frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y} + \lambda\left(\widehat{D}' - \widehat{D}'' \hat{c}_-\right)\right). \quad (32)$$

Remark 2. *When $g = \frac{1}{2}x^2$, $D(m) = \frac{1}{2}c^T K c$, resulting in $D' - D''\hat{c} = 0$. Consequently $\hat{c}$ can be written as a linear function of $\tilde{Y}$, simplifying the inferences considerably. Consequently a quadratic g enjoys both theoretical and practical advantages.*

To simplify exposition, we will focus on the quadratic case in this section. Define

$$\hat{c} = \left(\frac{1}{n}\hat{\beta}_1^2 \hat{Z}^T \hat{Z} + \lambda K\right)^{-1} \left(\frac{1}{n}\hat{\beta}_1 \hat{Z}^T \tilde{Y}\right). \quad (33)$$

This representation of c as a linear function of $\tilde{Y}$ allows us to use known results on linear smoothers for inferences.

Suppose that we use a p-degree truncated power series splines with M basis functions in our estimation. We make the following assumptions:

Assumption 1 $\{Y_i, X_i\}_{i=1}^n$ is an iid random sample satisfying

$$Y_i = \beta_0 + \beta_1 \int_0^{X_i} \left\{ exp\left(- \int_0^s (h(t))^2 dt \right) \right\} ds + u_i$$

where h is $p + 1$ times continuously differentiable and bounded on [0,1]. The error term is assumed to have mean zero and finite variance σ^2.

Assumption 2 The design points X_i are distributed according to density that is bounded above and away from zero, with a compact support on [0,1].

Assumption 3 The spline basis functions are pth degree truncated power series. The knots are equidistantly distributed on [0,1].

Assumption 4. The dimension of the spline basis functions grows with the sample size such that $M \sim C n^{1/(2p+3)}$, for some constant $C > 0$.

Assumption 5. The penalty parameter λ is assumed to grow with the sample size such that $\lambda = O(n^{\lambda})$ with $\lambda \leq 2/(2p + 3)$.

Remark 3. *Assumptions 2 and 3, standard in the spline literature, ensure that the design matrix is well behaved. In practice, the equidistant knots are sometimes replaced with equidquantile knots without affecting the asymptotic results. Assumption 4 indicates the growth rate of number of basis functions is identical to that for regression splines. Lastly*

under Assumption 5, the squared shrinkage bias (due to the penalization) is of smaller order than that of the variance and squared approximation bias, resulting in convergence rates analogous to the regression splines.

Let $\hat{h}(x) = h(x; \hat{c})$ be the solution to the penalized spline estimator (26) with a penalty term $D = \int_0^1 g(t)dt = \int_0^1 (h(t))^2 dt$. The consistency of the proposed estimator is established in the following theorem.

Theorem 1. *Under Assumptions 1-5, the average mean square error of the estimated* $\hat{h}$ *satisfies*

$$AMSE(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \hat{h}(X_i))^2 = O\left(n^{-\frac{2p+2}{2p+3}}\right).$$

Remark 4. *The average mean square error given above can be decomposed into the average mean square error of* $\hat{h}(X_i)$ *with respect to* $E[\hat{h}(X_i)]$ *and two additional terms involving the approximation error* $h(X_i) - E[\hat{h}(X_i)]$. *Kauermann et al. (2009) suggest that all three terms are of order* $n^{-(2p+2)/(2p+3)}$. *Since the exact form of the bias is generally unknown, below we present the confidence interval of* $\hat{h}(X_i)$ *around* $E[\hat{h}(X_i)]$.

Let W be a $n \times 2$ matrix with the ith row $W_i = (1, m(X_i)), i = 1, \dots, n$. Define

$$P_W = W(W^T W)^{-1} W^T,$$

$$P_Z = \left(\frac{1}{n}\beta_1 Z\right)\left(\frac{1}{n}\beta_1^2 Z^T Z + \lambda K\right)^{-1}\left(\frac{1}{n}\beta_1 Z^T\right). \quad (34)$$

Let, $\widehat{W}_i = \left(1, \widehat{m}(X_i)\right)$ and $\hat{P}_W$ and $\hat{P}_Z$ be defined similarly.

Under the assumption of iid errors, their variance σ^2 is estimated by the sum of squared residuals divided by proper degrees of freedom. Our semiparametric estimator has two parametric parameters β_0 and β_1, and a nonparametric smoother $m(X_i; \hat{c})$. The degrees of freedom of the smoother (or its equivalent number of coefficients to that of power series) is calculated as $tr\left(\hat{P}_Z\right)$. Therefore we estimate σ^2 with

$$s^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n - tr\left(\hat{P}_Z\right) - 2}.$$

Below we present an asymptotic variance estimator of the predicted values.

Theorem 2. Under Assumptions 1-5, the covariance matrix of the predicted values $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 m(X_i; \hat{c})$ satisfies

$$Var\left(\hat{Y}\right) = \sigma^2(P_W + \beta_1^2 P_Z^2 + 2\beta_1 P_W P_Z) + o_p(1) \equiv V + o_p(1). \quad (35)$$

Define $\hat{V} = s^2\left(\hat{P}_W + \hat{\beta}_1^2 \hat{P}_Z^2 + 2\hat{\beta}_1 \hat{P}_W \hat{P}_Z\right)$. $\hat{V} \xrightarrow{p} V$ as $n \to \infty$.

Denote by $\hat{V}_i$ the ith diagonal element of V. We construct the asymptotic $(1 - \alpha)\%$ variability band of $\hat{Y}_i$ by

$$\hat{Y}_i \pm z_{1-\alpha/2} s \sqrt{\hat{V}_i}, \quad (36)$$

where $z_{1-\alpha/2}$ is the corresponding critical value from the standard normal distribution.

Remark 5. *Alternatively, we can use the degrees of freedom of the residual in the calculation of variance. For linear smoothers, the residual degrees of freedom is given by* $2tr(\hat{P}_Z) - tr(\hat{P}_Z^2)$. *See, e.g., Ruppert et al. (2003) and references therein. In practice, these two specifications often give similar results.*

Remark 6. *The variability band (36) is about* $E[\hat{\beta}_0 + \hat{\beta}_1 m(X_i; \hat{c})]$ *rather than* $\beta_0 + \beta_1 m(X_i)$. *This is a well-known issue with series-based nonparametric estimations, of which the bias terms are generally not available. Although bias is inherent in nonparametric regression, approximate unbiasedness is often assumed and the variability band can be interpreted as approximate confidence interval. Since this approximate confidence interval is oftentimes over optimistic, Hastie and Tibshirani (1990) suggest replacing* $z_{1-\alpha/2}$ *in (36) with* $t_{1-\alpha/2,df}$, *where* df *is a proper degrees of freedom for nonparametric regressions. Eubank (1999) suggests Bonferroni methods to calculate confidence bands. Ruppert et al. (2003) discuss bias-corrected confidence intervals. Interested readers are referred to Hall and Opsomer (2005), Li and Ruppert (2008), Claeskens et al. (2009) and Kauermann et al. (2009) for recent developments on the asymptotic properties of penalized spline estimators.*

Remark 7. *Asymptotic normality can be established under additional assumption that entails undersmoothing for series based estimations. However the practical use of this result is limited since it rules out optimal smoothing parameters selected according to the CV/GCV criterion.*

Remark 8. *Our estimator is semiparametric with two parametric coefficients. Taking* $\hat{m}$

as nuisance parameters, the estimator can be viewed as a two-step estimator with nonparametric first step estimates. Newey (1994) and Ai and Chen (2007) discuss the estimation of asymptotic semiparametric variance of the second stage estimates. Recently Acerberg et al. (2011) show that the asymptotic parametric variance that ignores the nonparametric nature of the first stage (for instance, the method of Newey (1984)) is numerically identical to the semiparametric variance. In particular, Acerberg et al. (2011) provide several examples that use sieve estimators in the first step. The penalized spline estimator investigated in this study fits into this framework naturally.

Remark 9. *Although we only consider the quadratic case (33), we find in our numerical experiments that (36) provides a reasonable approximate variability bands for estimates from the more general case (32).*

Remark 10. *We present the alternative representation (32) to facilitate the asymptotic analysis. Our numerical experiments indicate that the Gauss-Jordan algorithm given in the previous section is usually more robust and converges faster, especially when a non-qudratic g is used. We recommend the Gauss-Jordan algorithm for the implementation of our estimator.*

3.5. Specification of Spline Basis and Smoothing Parameter

Implementation of the penalized spline estimators entails the specification of the spline basis and smoothing parameters. The former includes the type of splines, number and location of knots. Commonly used splines include the truncated power series, B-splines

and radial basis splines. The spline literature indicates that the practical differences among these splines are oftentimes quite small. For simplicity, we use the third degree truncated power series basis functions (the cubic splines).

Because the penalized spline estimations normally use a relatively generous spline basis, the number and location of knots plays a relatively minor role in the estimates. In this study, we follow the auto knot selection rule in Ruppert (2002), where the number of knots is given by

$$M = \min\left(\frac{1}{4} \times number\ of\ unique\ X_i, 35\right),$$

and the knots are placed at the $(m+1)/(M+)$th sample quantile of the unique X_i's for $m = 1, \ldots, M$.

It is well known in the spline literature that the estimation results generally depend crucially on the smoothing parameter, but to a large degree are not sensitive to the specification of other aspects (See e.g., Ruppert, 2002). A commonly used approach of smoothing parameter selection is the principle of cross validation. Let $\hat{Y}_{(i)}$ be the estimate of Y_i from a given estimator that uses all but the ith observations. The 'leave-one-out' cross validation criterion, in terms of sum of squared residuals, is given by

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_{(i)}\right)^2.$$

Direct implementation of the cross validation is straightforward but often costly,

especially for estimators without analytical solutions. For linear estimators, there exist exact formulae to evaluate the cross validation criterion function, using regression results based on the full sample alone. For nonlinear estimations, this exact solution usually does not exist. Nontheless, there exist approximate formulations that have been shown to give rather close results.

In this section, we derive an approximate formula of the cross validation criterion function for the proposed estimator. For each $i = 1, \dots, n$, denote $\hat{c}_{(i)}$ be the solution to

$$\sum_{k=1,k\neq i}^{n} (Y_k - \beta_0 - \beta_1 m(X_k; c))^2 + \lambda D\big(f(x)\big).$$

We establish the following result.

Theorem 3. *Let s_i be the ith diagonal element of $\hat{P}_Z$ given in (34), $i = 1, \dots, n$. The Cross Validation (CV) criterion satisfies*

$$CV = \sum_{i=1}^{n}\big(Y_i - \hat{Y}_{(i)}\big)^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - s_i}\right)^2 + o_p(1). \tag{37}$$

The Generalized Cross Validation (GCV) is a popular alternative to the CV criterion . The GCV is obtained by replacing $1 - s_i$ in (37) with $1 - \frac{1}{n} tr\big(\hat{P}_Z\big)$ (see, e.g., Wahba 1990). One can readily infer from Theorem 3 that in our case

$$CGV \approx \sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - \frac{1}{n} tr\big(\hat{P}_Z\big)}\right).$$

Remark 11. *An alternative criterion of smoothing parameter selection is the estimated risk (see, e.g., Eubank 1999). Although conceptually simple, this criterion entails a proper prior estimate of* σ^2*, which complicates the issue since optimal smoothing parameters for conditional mean estimations generally are not optimal for variance estimations.*

3.6. Multiple Regressions

In this section we consider the case where $y(\cdot)$ is a function of J variables, being monotone and concave in each one. With a slight abuse of notation, we use the same notations as in the single covariate case, adding additional subscripts to index covariates whenever needed. For simplicity, we focus on the case of general additive models:

$$Y_i = \beta_0 + \sum_{j=1}^{J} \beta_j m_j\left(X_{j,i}\right) + u_i.$$

For general treatments of additive models, see Hastie and Tibshirani (1990).

We estimate the additive model using the penalized spline estimator by minimizing the following objective function:

$$\frac{1}{n}\left(\sum_{i=1}^{n} Y_i - \beta_0 - \sum_{j=1}^{J} \beta_j m_j\left(X_{j,i}\right)\right)^2 + \sum_{j=1}^{J} \lambda_j D\left(m_j(x)\right).$$

The Gauss-Jordan algorithm described above for the single covariate case can be

extended readily to the multiple covariates case by updating the coefficients c_j, $j = 1, \dots, J$ sequentially via back-fitting. Alternatively, we can update all coefficients simultaneously to improve efficiency. For $j, k \in (1, \dots, J)$, let

$$\hat{R}_{j,k} = \begin{cases} \frac{1}{n}\hat{\beta}_j^2 \hat{Z}_j^T \hat{Z}_j + \lambda_j D_j'', & if\ j = k; \\ \frac{1}{n}\hat{\beta}_j \hat{\beta}_k \hat{Z}_j^T \hat{Z}_k, & if\ j \neq k, \end{cases}$$

where $\hat{Z}_j = (\hat{Z}_{j,1}, \dots, \hat{Z}_{j,n})$ with $\hat{Z}_{j,i} = dm_j(X_{j,i}; \hat{c}_j)/dc_j$ for $j = 1, \dots, J$ and $i = 1, \dots, n$.

Further define $\hat{c} = (\hat{c}_1^T, \dots, \hat{c}_J^T)^T$, $\hat{S} = (\hat{S}_1^T, \dots, \hat{S}_J^T)^T$, and

$$\hat{R} = \begin{bmatrix} \hat{R}_{1,1} & \cdots & \hat{R}_{1,J} \\ \vdots & \ddots & \vdots \\ \hat{R}_{J,1} & \cdots & \hat{R}_{J,J} \end{bmatrix}.$$

The coefficints $\hat{c}$ are then updated according to

$$\hat{c} = \hat{c}_- - \hat{R}^{-1}\hat{S}. \tag{38}$$

Note that if we set all the off-diagonal blocks of $\hat{R}$ to null matrices (i.e., $\hat{R}_{j,k} = 0$ for all $j \neq k$), updating by (38) is equivalent to updating each component sequentially and generally less efficient.

The methods for knots and smoothing parameter selection and construction of confidence intervals discussed in the previous sections can be generalized to the multiple regression case in a straightforward manner.

3.7. Monte Carlo Simulations

In this section we provide numerical evidence on the performance of the proposed estimator. We consider the following data generating process:

$$Y_i = \beta_0 + \beta_1 \log X_{1,i} + \beta_2 \log X_{2,i} + u_i$$

where $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = -2$, $u_i \sim N(0,1/10)$ and $X_{1,i}$ and $X_{2,i}$ are generated randomly according to the Gamma distributions $\Gamma(1,2)$ and $\Gamma(2,1)$ respectively.

We use the default knot selection method and GCV method for smoothing parameter selection. For comparison, we also consider the following simple polynomial model

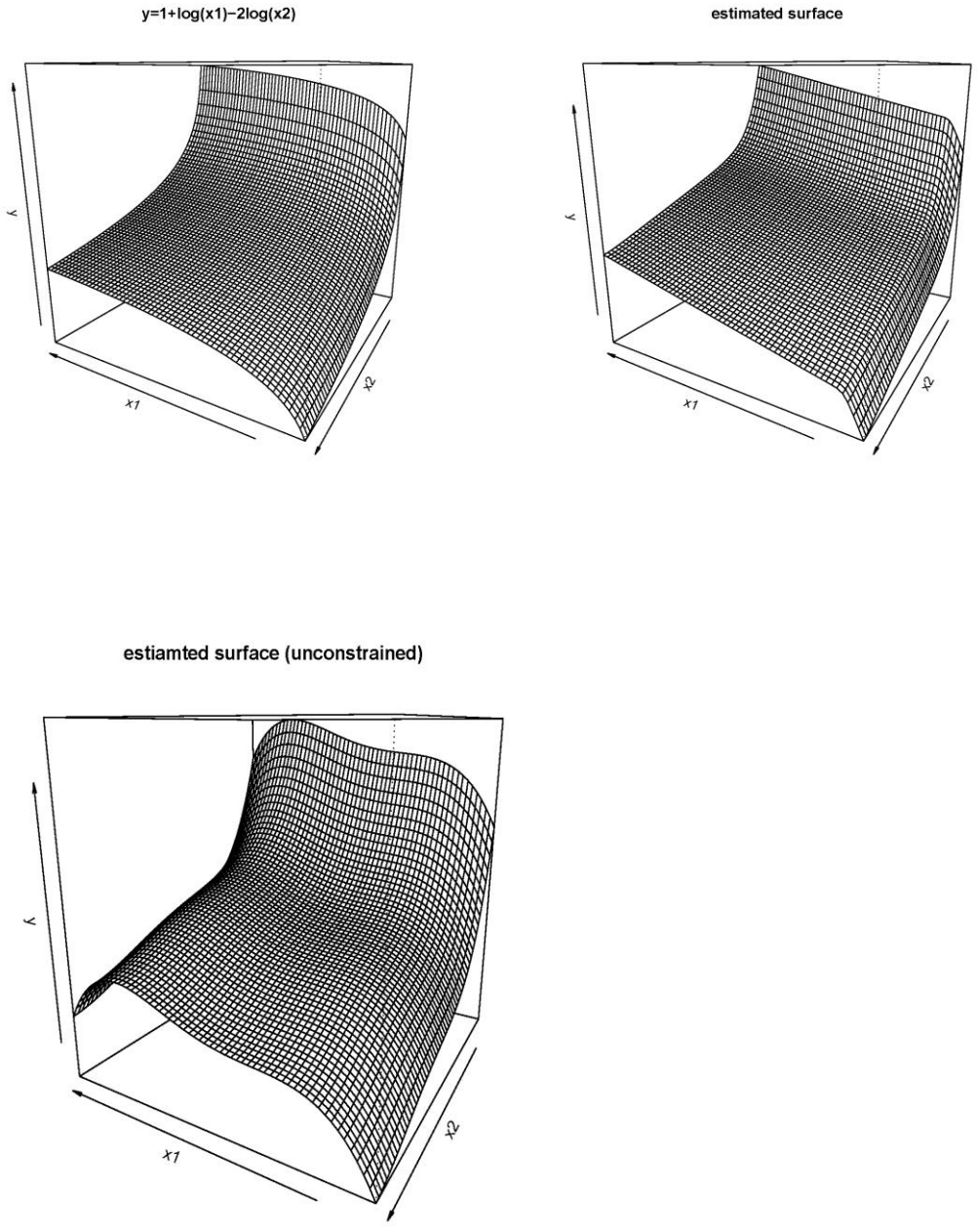
$$f(X_i) = \beta_0 + \sum_{k=1}^{4} (\beta_{1,k} X_{1,i}^k + \beta_{2,k} X_{2,i}^k) \quad (39)$$

We set the sample size $n = 100$ and number of repetitions 100 as well. All experiments on the proposed method converge successfully. Below we first provide an illustration of the fitted surface. In Figure 17, the true surface and the estimated constrained surface are reported in the top panel, followed by the estimated unconstrained surface by the polynomial model below. Although both estimates capture the overall shape of the underlying relationship, the unconstrained estimate clearly violates monotonicity and concavity and is subject to the oscillations associated with high order polynomials.

Next we present some numerical results of the simulations. The average and

median estimated mean squared errors for the constrained model are 0.102 and 0.125 respectively. Their counterparts for the unconstrained model are 0.132 and 0.162.

Figure 17.True and Estimaded Surfaces.



We also evaluate the goodness-of-fit by assessing the distance between the true surface

and the estimated surface. For a given vector (X_1, X_2), we calculate $d_1 = |(f(X_1, X_2) - \hat{f}(X_1, X_2)|$ and $d_2 = (f(X_1, X_2) - \hat{f}(X_1, X_2))^2$, where f stands for the true data generating process, and $\hat{f}$ is either of the two estimated models. We evaluate both distances on a equally spaced grid of $X_1 \in [0,3]$ and $X_2 \in [0,6]$ with 500 increments in either dimension. The medians of the average distance d_1 and d_2 for the constrained model are 0.483 and 0.683 respectively. The corresponding values for the unconstrained model are 1.989 and 0.772.[20]

Table 17: Summary statistics of production data

	Mean	S.D.	Min	Max
Output	16.3	8.3	1.7	37.1
Capital	4.8	2.8	9.6	0.3
Labor	57.7	27.2	1.1	98.9

Our simulations offer some support to the proposed model. It clearly preserves the desired monotonicity and concavity. In addition, the built-in shape restrictions to a large degree suppress the oscillations usually associated series-based nonparametric estimations. This is particularly valuable when high order series are used in estimations.

3.8. Empirical Applications

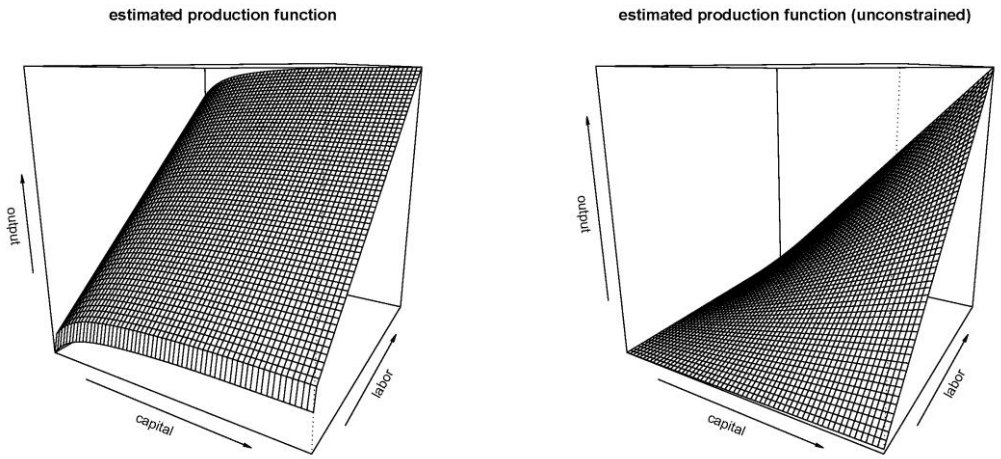
In this section, we present an illustrative application of the proposed model to the empirical estimation. Here we will use two datasets, one originally used in Coelli (1996) containing information to estimate a prodcution function, while the second set comes

[20] We do not report the averages of d_1 and d_2 for the two models. This comparison clearly favors the constrained model because the oscillations of the linear polynomial model lead to some exceptionally large values near the boundaries of the evaluation grid, especially for d_2.

from the Dyer, Kagel and Levin (1989) paper, later used by Bajari and Hortaçsu (2005) and kindly provided to us by Christopher Parmeter, which contains information on the first-price sealed bid auction.

First data set contains information on the level of output, capital and labor inputs for a cross-section of 60 firms. Table 1 reports some summary statistics of the data set.

Figure 18. Constrained and Unconstrained Surfaces of the Production Function

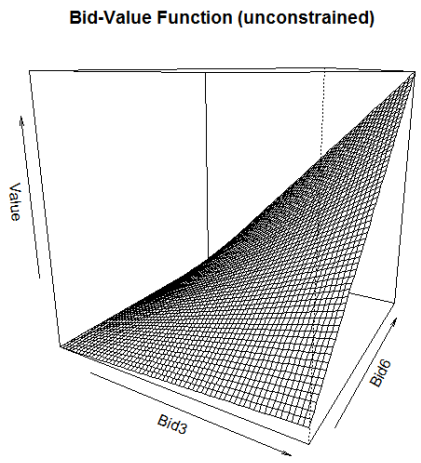


As in the previous section, we estimate both the proposed constrained model and the unconstrained linear polynomial model. The estimated surface for the production function are reported in Figure 18, with the constrained estimate on the left and the unconstrained one on the right. Both estimates indicate positive marginal contributions of the inputs. However the constrained estimate demonstrates concavity, while the unconstrained clearly violates it.

Second dataset contains values and bids provided by the participants of a sealed

bid auction with either three or six bidders. Twenty three auctions over three experimental runs are considered here, with total of 414 bids. Estimation of constraind models for both number of bidders returned a result which didn't converge. To the contrary, models with unconstrained setup returned a result where an auction with three and an auction with six bidders both presented concavity in a bid-value relationship. We also note that with larger number of bidders bids closer follow the value of the bidder in the auction.

Figure 19. Bid - Value Function



3.9. Concluding Remarks

Flexibility of functional forms and adherence to theory are main considerations in economic modelling. Monotonicity and concavity are the primary requirements for a good model. Oftentimes models satisfy one of the two requirements and sometimes may

have them at odds with each other. Building on previous studies we propose a combination of transformation and semiparametric estimation procedures which may help in producing plausible results.

By parametrization of the model we provide an esimator which is easier to implement, especially in the multiple regression setting, in our case is the penalized spline estimator, which helps to balance the fidelity to the data.

In our paper based on cetrtain assumptions on the sample, spline basis and penalty paramenter we derive the average mean square error and the covarance matrix of the predicted variales. We aslo derive the cross validation criterion for the proposed procedure.

Simulated data allowed us to compare the estimated and actual surfaces of the production function and showed that our method captured both monotonicicty and concavity requirements. Application to the real data on the firm level as well as individuals showed that and unconstrained estimation captures the concavity while the unconstrained does not.

References

Ai C, Chen X (2007) Estimating of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. Journal of Econometrics 141: 5-43.

Aerts M, Claeskens G, Wand M (2002) Some theory for penalized spline additive models. Journal of Statistical Planning and Inferences 103: 455-470.

Ackerberg D, Chen X, Hahn J (2011) A practical asymptotic variance estimator for two-step semiparametric estimators. Working paper.

Andrews, D.F. and Pregibon D. (1978), Finding the Outliers that Matter, Journal of the Royal Statistical Society, Series B (Methodological), 40(1), pp. 85-93

Apt, J. (2005), Competition Has Not Lowered US Industrial Electricity Prices, The Electricity Journal, 18(2), pp. 52-61.

Axelrod, H.J., DeRamus, D.W. and Cain, C. (2006), The Fallacy of High Prices, Public Utilities Fortnightly, 144(11), pp. 55-60.

Bădin L., Daraio C. and L. Simar (2010). Optimal Bandwidth Selection for Conditional Efficiency Measures: A Data-driven Approach, European Journal of Operations Research, vol. 201(2), pp. 633-640.

Bajari P. and A. Hortaçsu (2005). Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data. Journal of Political Economy, vol. 113, pp. 703-741.

Banker R.D. and R. Morey (1986). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. Operations Research, vol. 34(4), pp. 513-521.

Braun WJ, Hall P (2001) Data sharpening for nonparametric inference subject to constraints. Journal of Computational and Graphical Statistics 10: 786806.

Brunk HD (1955) Maximum likelihood estimates of monotone parameters. Annals of Mathematical Statistics 26: 607616.

Caudill S.B. and J. Ford (1993). Biases in Frontier Estimation due to Heteroskedasticity, Economics Letters, vol. 41(1), pp. 17-20.

Cazals C., Florens J.P. and L. Simar (2002). Nonparametric Frontier Estimation: a Robust Approach, Journal of Econometrics, 106, pp. 1-25.

Charnes, A., Cooper, W. W. and Rhodes, E. (1978), Measuring Efficiency of the Decision Making Units, European Journal of Operational Research, 2(6), pp. 429-444.

Chambers R (1988) Applied Production Analysis: A Dual Approach. Cambridge University Press.

Chernozhukov V, Fernandez-Val I, Galichon A (2007) Improving estimates of monotone functions by rearrangement. Mimeo.

Christensen, L.R. and Greene, W.H. (1976), Economies of Scale in U.S. Electric Power Generation, Journal of Political Economy, 84(4), pp. 655-676.

Coelli, TJ (1996) A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation. CEPA Working Paper 96/7, Department of Econometrics, University of New England, Armidale NSW Australia.

Daraio C. and L. Simar (2005). Introducing Environmental Variables in Nonparametric Frontier Models: a Probabilistic Approach, Journal of Productivity Analysis, vol. 24 (1), pp. 93-121.

Daraio C. and L. Simar (2007). Conditional Nonparametric Frontier Models for Convex and Non-convex Technologies: a Unifying Approach, Journal of Productivity Analysis, vol. 28(1-2), pp. 13-32.

De Boor C (2001) A practical guide to splines. Springer.

Demchuk P. and V. Zelenyuk (2009).Testing Differences in Efficiency of Regions Within a Country: the Case of Ukraine, Journal of Productivity Analysis, vol. 32, pp. 81-102.

Dyer D., Kagel, JH and D. Levin (1989). Resolving Uncertainty about the Number of Bidders in Independent Private-Value Auctions: An Experimental Analysis. RAND Journal of Economics, vol.20, pp. 268-279.

Electricity Retail Energy Deregulation Index 2003 (2003), Center for the Advancement of Energy Markets, Retrieved June 8, 2011, from www.caem.org/Content/RED%20Index/RED_Index_2003.htm

Eubank LE (1999) Nonparametric regression and simple smoothing. Marcel Dekker.

Färe, R., Grosskopf, S., Lindgren, B. and Roos, P. (1992), Productivity Changes in Swedish Pharamacies 1980–1989: A non-parametric Malmquist approach, Journal of Productivity Analysis, 3(1-2), pp. 85-101.

Farell, M. J. (1957), The Measurement of Productive Efficiency. Journal of the Royal Statistical Society, 120 (Series A General, Part III), pp. 253-281.

Gocht A. and K. Balcombe (2006). Ranking Efficiency Units in DEA Using Bootstrapping an Applied Analysis for Slovenian Farm Data, Agricultural Economics, vol. 35, pp. 223-229.

Hall P, Huang H (2001) Nonparametric kernel regression subject to monotonicity constraints. The Annals of Statistics 29(3): 624647.

Hall P, Opsomer JD (2005) Theory for penalized spline regression. Biometrika, 92: 105-118.

Härdle W, Sylvie H, Mammen E, Sperlich S (2004) Bootstrap inference in semiparametric generalized additive models. Econometric Theory 20(2): 265-300.

Hastie, TJ, Tibshirani, R J (1990). Generalized additive models. London: Chapman & Hall.

Heckman, J., Ichimura, H. and Todd, P., (1997), Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Program, Review of Economic Studies, 64(4), pp. 605-654

Hall P., Racine J.S. and Q. Li (2004). Cross-validation and the Estimation of Conditional Probability Densities, Journal of the American Statistical Association, vol. 99(486), pp. 1015-1026.

Joskow, P.L. (2006), Markets for Power in the U.S.: An Interim Assessment, The Energy Journal, 27(1): 1-36.

Kauermann G, Krivobokova T, Fahrmeir L (2009) Some asymptotic results on generalized penalized spline smoothing. Journal of the Royal Statitical Society, Series B 71, 487-503.

Kelly C, Rice J (1990) Monotone smoothing with application to dose response curves and the assessment of synergism. Biometrics 46: 10711085.

Keramidou I., Mimis A. and Pappa E. (2010).Determinants of Efficiency of Prepared Meat Products Industryin Greece, European Journal of Social Sciences, vol. 17(4), pp. 509-520.

Kiesling, L.L. and Kleit, A.N. (Eds.) (2009), Electricity Restructuring: The Texas Story, AEI Press

Kneip A, Park B, Simar L (1998) A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores, Econometric Theory, vol. 14(6), pp.783–793.

Lechner, M (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Econometric Evaluation of Labour Market Policies, ed. Michael Lechner and Friedhelm Pfeiffer, pp. 43-58.

Li Q, Racine JS (2007) Nonparametric econometrics: Theory and practice. Princeton University Press.

Li Y, Ruppert D (2008) On the asymptotics of penalized splines. Biometrika 95: 415-436.

Löber G. and M. Staat (2010).Integrating Categorical Variables in Data Envelopment Analysis Models: A simple Solution Technique, European Journal of Operational Research, vol. 202(3), pp. 810-818.

Mansur, E.T. and White, W.W. (2009), Market Organization and Efficiency in Electricity Markets, Discussion Draft, Retrieved June 8, 2011, from http://bpp.wharton.upenn.edu/mawhite/papers/MarketOrg.pdf

Mammen E (1991) Estimating a smooth monotone regression function. Annals of Statistics 19(2): 724740.

Mammen E, Thomas-Agnam C (1999) Smoothing splines and shape restrictions. Scandinavian Journal of Statistics 26: 239252.

Mayen C., Balagtas J. and Alexander C. (2010).Technology Adoption and Technical Efficiency: Organic and Conventional Dairy Farms in the United States, American Journal of Agricultural Economics, vol. 92(1), pp.181-195.

Mukerjee H (1988) Monotone nonparametric regression. Annals of Statistics 16: 741750.

Muñiz M. (2002). Separating Managerial Inefficiency and External Conditions in Data Envelopment Analysis, European Journal of Operational Research, vol. 143(3), pp. 625-643.

Nadaraya E.A. (1964). On estimating regression, Theory of Probability Applications, vol. 9(1), pp. 141-142.

Newey WK (1984) A method of moment interpretation of sequential estimators. Economic Letters 14: 201-206.

Newey WK (1994) The asymptotic variance of semiparametric estimators. Econometrica 62: 1349-1382.

Nishimizu, M. and Page, J. M. (1982), Total Factor Productivity Growth, Technological Progress and Technical Efficiency Change: Dimensions, The Economic Journal, 92(368), pp. 920-936.

Racine JS, Parmeter, C (2010) Constrained Nonparametric Kernel Regression: Estimation and Inference. Mimeo.

Ramsay JO (1988) Monotone regression splines in action (with comments). Statistical Science 3: 425461.

Ramsay JO (1998) Estimating smooth monotone functions. Journal of the Royal Satistical Society Series B 60 (2): 365-375.

Rose, K. (2007), The Impact of Fuel Costs on Electric Power Prices. Working Paper. Retrieved June 8, 2011, from http://www.kenrose.us/sitebuildercontent/sitebuilderfiles/impactoffuelcostsonelectricpowerprices_final.pdf

Rose, K. and Meeusen, K. (2005), Performance Review of Electric Power Markets: Update and Perspective, Report for the Virginia State Corporation Commission.

Rosenbaum, P.R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects, Biometrica vol. 70(1), pp. 41–55.

Ruggiero J. (1996). On the Measurement of Technical Efficiency in the Public Sector, European Journal of Operational Research, vol. 90(3), pp. 553-565.

Ruggiero J. (1998). Non-discretionary Inputs in Data Envelopment Analysis, European Journal of Operational Research, vol. 111(3), pp. 461-469.

Rungsuriyawiboon, S. and Coelli, T. (2006), Regulatory Reform and Economic Performance in US Electricity Generation, in Performance Measurement and Regulation of Network Utilities, Coelli T. and Lawrence D. eds., pp. 267-292.

Ruppert D (2002) Selecting the number of knots for penalized splines. Journal of Computational Statistics 11: 735-757.

Ruppert D, Wand W, Carroll R (2003) Semiparametric regression. Cambirdge University Press.

Simar, L. (2003), Detecting Outliers in Frontier Models: A Simple Approach, Journal of Productivity Analysis, 20(3), pp. 391-424.

Simar L. and P. Wilson (2007). Estimation and Inference in Two-stage, Semi-parametric Models of Production Processes, Journal of Econometrics, vol. 136(1), pp. 31–64.

Simar, L. and Zelenyuk, V. (2007), Statistical Inference for Aggregates of Farrell-Type Efficiencies, Journal of Applied Econometrics, 22(7), pp. 1367–1394.

Sioshansi, F.P. and Pfaffenberger, W. (Eds.). (2006), Electricity Market Reform: An International Perspective, Elsevier.

Syrjänen M. (2004). Non-discretionary and Discretionary Factors and Scale in Data Envelopment Analysis, European Journal of Operational Research, vol. 158(1), pp. 20-33.

Texas Senate Bill 7, Retrieved June 8, 2011, from http://www.capitol.state.tx.us/BillLookup/Text.aspx?LegSess=76R&Bill=SB7

Yang Z. and J. C. Paradi (2006). Cross Firm Bank Branch Benchmarking Using "Handicapped" Data Envelopment Analysis to Adjust for Corporate Strategic Effects, Proceedings of the 39th Hawaii International Conference on System Sciences.

Wahba G (1990) Spline models for observational data. SIAM.

Wand M (1999) On the optimal amount of smoothing in penalized spline regression. Biometrika 86: 936-940.

Wang H. and P. Schmidt (2002). One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels, Journal of Productivity Analysis, vol. 18(2), pp. 129-144.

Watson G.S. (1964). Smooth Regression Analysis, Sankhya Series A, 26, pp. 359-372.

Wilson, P.W. (1993), Detecting Outliers in Deterministic Nonparametric Frontier Models with Multiple Outputs, Journal of Business and Economic Statistics, 11(3), pp. 319-323.

Wu, X. and R. Sickles (2012), Semiparametric Estimations with Shape Constraints, mimeo, Rice University.

Wu, X., Sickles, R. and P. Demchuk (2012), Extensions on the Semiparametric Estimations with Shape Constraints of Wu and Sickles, mimeo, Rice University.

Zarnikau, J., Fox, M. and Smolen, P. (2007), Trends in prices to commercial energy consumers in the competitive Texas electricity market, Energy Policy, 35(8), pp. 4332-4339

Zarnikau, J. and Whitworth, D. (2006), Has Electric Utility Restructuring Led to Lower Electricity Prices for Residential Consumers in Texas?, Energy Policy, 34(15), pp. 2191-2200.